FACULTAD DE ESTUDIOS ESTADÍSTICOS

Máster de Minería de Datos e Inteligencia de Negocios

Trabajo de Fin de Máster

Metodología de Minería de Datos Aplicada a la Competencia Hotelera

Alumna: Inmaculada Gutiérrez García-Pardo
Tutor: Javier Castro Cantalejo

Noviembre 2016

UNIVERSIDAD COMPLUTENSE
MADRID

A mi familia, no puedo escribir tantos motivos
A Javier Castro, por su ayuda imprescindible
A Dani, por tener tiempo, paciencia y respuesta para todas mis preguntas
A Aarón, que ha vivido conmigo este proyecto

Resumen

El presente estudio comprende una síntesis general de los conocimientos adquiridos en el desarrollo del Máster en Minería de Datos e Inteligencia de Negocios. Se ha intentado pasar por la mayoría de las áreas que en el mismo se tratan, prestando especial atención a la parte de análisis de datos propiamente dicha.

La temática se ha centrado en el sector hotelero de la ciudad de Madrid. Se pretende hacer un ejercicio en profundidad de análisis de datos, seguido de un análisis de predicción del precio de los hoteles situados en esta ciudad, tomando como referencias distintas características de estos establecimientos, además de momentos temporales y otros matices. Otro punto a tratar en este estudio está basado en un análisis de la competencia hotelera, que tomará como base los resultados obtenidos en los primeros pasos de este proyecto.

Así, se llega a la selección de un modelo óptimo de predicción, obtenido tras un proceso de ensayo-error de distintas técnicas predictivas, seguido de un proceso de elección. Así mismo, se consigue entender cómo se agrupan los distintos hoteles y cómo se sitúan en su mercado, atendiendo al comportamiento de los centros que forman su competencia.

Abstract

This study comprises a general synthesis of the knowledge acquired in the development of the Master Degree Data Mining and Business Intelligence. It has been trying to get through most of the treated areas, with special attention in data analysis.

The thematic has focused on the hotelier sector of Madrid. It aims to do a deep exercise in data analysis, followed by an analysis of price prediction about hotels located in the city, taking as references different characteristics of these establishments, in addition to temporary moments and nuances. Another point to be discussed in this study is based on an analysis of hotelier competition, which will build upon the results obtained in the first steps of this project.

Thereby, we arrive at the selection of an optimal prediction model, obtained after a trial-error process of different predictive techniques, followed by a process of choice. Likewise, it is possible to understand how the different hotels are grouped and how they placed in their market, attending to the behavior of the centers that form their competence.

Índice

1. Naturaleza de los Datos

Booking es una empresa subsidiaria de *The Princeline Group*¸ que opera con su marca *Booking.com*, y es líder mundial en reservas de alojamiento online. Este sitio web y sus aplicaciones reciben visitantes en busca de estancias de ocio y negocios a nivel internacional. Cada día se reservan más de 1.100.000 a través de él. Esta empresa está en funcionamiento desde 1996, y su sitio web está disponible en más de 40 idiomas, con una oferta de 1.042.512 alojamientos activos en 227 países y territorios.

Los datos objeto del análisis en el presente estudio recogen información sobre los 278 hoteles madrileños registrados en el sitio web *Booking* durante el momento de la obtención de la información. Los datos recogidos abarcan tanto aspectos descriptivos de los distintos hoteles, como lo son su nombre, número de estrellas, código postal y distrito financiero en que se encuentra, como las valoraciones recibidas por parte de los usuarios (estas valoraciones puntúan, entre otros aspectos, los distintos servicios e instalaciones ofrecidos por el hotel en cuestión). También recogemos la fecha de inscripción de cada hotel en el sitio web, y la cantidad de comentarios dejados por los usuarios, variables que nos permitirán conocer y jugar con la popularidad de cada centro. Por último, se incluye también el precio publicado por cada hotel para una habitación doble simple, para cada una de las 31 noches del mes de Agosto de 2016, además de la indicación de si dicho precio se corresponde con una oferta ofrecida para una noche en concreto, o si, por el contrario, es el precio estándar marcado. Así, contaremos con aproximadamente 25.000 datos.

Toda la información está recogida en un mismo momento temporal, con el objetivo de obtener una visión a futuro de la variación de los precios. Así, hemos podido observar, por ejemplo, como fluctúan las distintas ofertas ofrecidas, comparando si la búsqueda se realiza con una semana de margen con el día de la reserva, o a mes vista.

Se ha focalizado la atención del estudio en la ciudad de Madrid, ya que la autora del estudio reside en dicha ciudad, y se asume que el conocimiento de la ciudad aportará un extra de información a la hora de entender los datos, utilizarlos en los distintos procesos que a continuación se explicarán, y obtener las conclusiones oportunas.
Cabe mencionar, que una vez obtenidos los datos de partida, se precisó de un amplio proceso de depuración, en el que se buscó la unificación de los formatos, así como la organización y estandarización de los mismos, de forma que la información contenida pudiera ser aprovechada al máximo, y vista e interpretada desde diferentes puntos de vista con el objetivo de darle diferentes sentidos, atendiendo a la necesidad de cada situación.

2. Objetivos y metodología

El primer paso en la elaboración del estudio es establecer los objetivos fundamentales del mismo, así como la metodología necesaria para su elaboración.
Así, el objetivo principal perseguido se resume en las siguientes líneas:

- Conocimiento, estudio y análisis del comportamiento de los hoteles de Madrid. Buscaremos comprender la relación del precio de los distintos establecimientos Con respecto a las distintas características físicas de cada hotel, como puedan ser la zona en que se encuentra, su número de estrellas o ubicación, así como de otros aspectos, como el momento temporal escogido para la reserva o la opinión de los clientes. En este punto se intentará entender cómo los matices mencionados afectan ante la decisión de un hotel a la hora de establecer un precio. Conocer estas relaciones será fundamental para llegar a comprender el comportamiento de los distintos hoteles, además de para intentar conocer los distintos precios que previsiblemente serán fijados para meses venideros. Se tendrán en cuenta conceptos tan importantes como la "competencia" de un hotel, puesto que es importante entender que un establecimiento marca sus tarifas atendiendo no sólo a sus características individuales, sino también a la del resto de empresas, cuyo comportamiento le podría llevar a una fluctuación en el número de posibles clientes. Esta consideración nos posibilitará llevar a cabo un análisis de competencia empresarial hotelera, lo cual nos permitirá añadir matices al estudio del comportamiento de los hoteles que nos ocupan.

Con objeto de alcanzar este punto, nos planteamos unos objetivos secundarios, que servirán de guía a la hora de alcanzar las conclusiones finales. Estos objetivos se presentan resumidos a continuación:

- Estudio de los datos y variables en uso.
- Análisis de las correlaciones entre variables.
- Análisis predictivo del precio de los distintos hoteles. Conociendo todos los precios de los distintos centros para un momento temporal concreto, además de distintas características físicas de los establecimientos y diferentes datos de interés proporcionados por *Booking*, este punto nos permitirá responder a la cuestión ¿Es posible predecir el precio de un hotel para cualquier momento del año.
- Agrupación de variables en grupos homogéneos y distintos entre sí.
- Fijación del concepto *competencia*, atendiendo a distintas situaciones.
- Estudio del comportamiento de un hotel concreto, en comparación con su competencia.

Para cumplir los objetivos establecidos, se marcan unos pasos a seguir, que nos conducirán a los puntos finales del estudio que nos ocupa. Algunos de los más importantes son:

- Conocer a fondo toda la información obtenida a partir de un profundo análisis descriptivo.
- Establecer una metodología clara sobre las técnicas de predicción utilizadas.

- Interpretación, comprensión y simplificación de las distintas variables en uso, así como de los datos a tratar. Para ello, nos apoyaremos en distintas técnicas de análisis factorial y *cluster*.
- Dividir la base de datos en tres partes (*Training, Validation* y *Test,* o, lo que es lo mismo, Aprendizaje, Validación y Test). Esto nos permitirá conseguir modelos más óptimos, en cuando a validez predictiva se refiere.
- Encontrar los mejores parámetros estructurales a utilizar en las diferentes técnicas de predicción utilizadas.
- Establecer una comparativa que permita señalar el mejor modelo predictivo obtenido.
- Realizar un post-análisis que ayude a entender el comportamiento del modelo seleccionado como mejor modelo de predicción.
- Estudio del comportamiento de distintos grupos de hoteles similares entre sí.

Las técnicas estadísticas multivariantes (Peña, 2002) que se van a utilizar para llevar a cabo este estudio se detallan a continuación:

- **Análisis Factorial** esta técnica se utilizará para agrupar y simplificar las variables.
- **Análisis *Cluster*,** técnica que nos permitirá clasificar los hoteles en uso, atendiendo a las semejanzas y diferencias de perfiles existentes entre los comportamientos de la población.
- **Regresión Lineal, Redes neuronales, *Random Forest* y*Gradient Boosting*,** técnicas utilizadas para la predicción y obtención del mejor modelo.

Para el desarrollo del estudio se ha buscado apoyo en la metodología SEMMA(Rodríguez & al, 2003), de la cual se presenta un resumen a continuación. También se dará una explicación breve de cada una de las técnicas estadísticas mencionadas y de cómo compararlas, así como del software utilizado para llevar a cabo todo el proceso.

2.1. Metodología SEMMA

Para alcanzar los objetivos establecidos, se ha intentado seguir el patrón de trabajo establecido por la Metodología SEMMA(*Sample, Explore, Modify, Model Asses,* o, lo que es lo mismo, Muestrear, Explorar, Modificar, Modelizar, Evaluar).

Este tipo de metodología se estructura como se muestra en el siguiente diagrama[1].

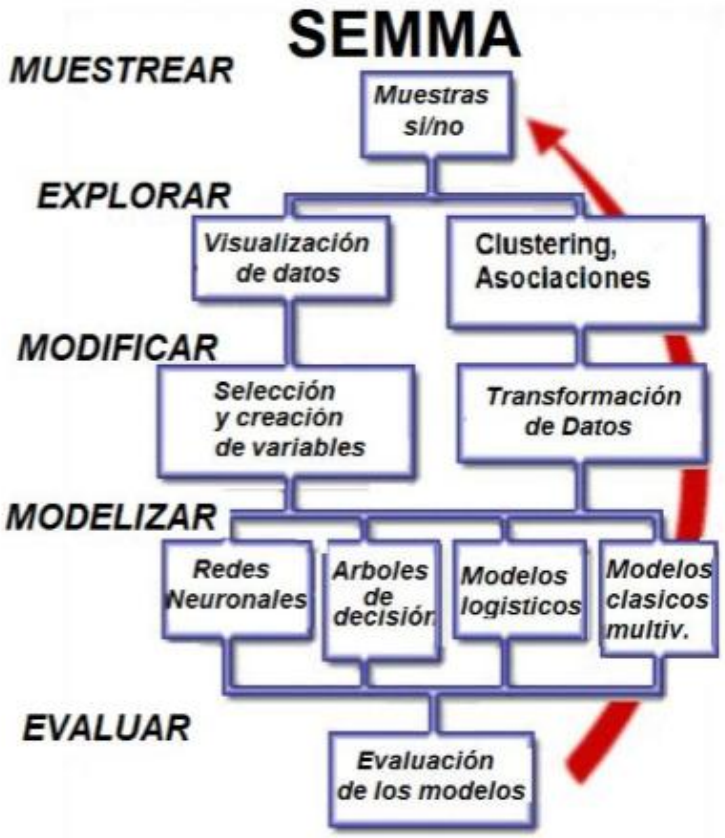

Figura 1. Diagrama SEMMA

Aunque este es el patrón a seguir propuesto inicialmente, este esquema no ha de seguirse exhaustivamente, ni en orden ni en contenido, puesto que no siempre intervienen todas las fases que aparecen en el esquema, a menudo el orden no es exacto, y en ocasiones, este proceso puede llegar a repetirse muchas veces, pasando de unas fases a otras sin respetar el orden del flujo.

2.2. Análisis Factorial

El Análisis Factorial (En línea 2016 a) es una técnica estadística de reducción de datos usada para explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas, llamadas factores. Para ello utiliza un conjunto de variables aleatorias inobservables, que llamaremos factores comunes, de forma que todas las covarianzas o correlaciones son explicadas por dichos factores y cualquier porción de la varianza inexplicada por los factores comunes se asigna a términos de error residuales que llamaremos factores únicos o específicos. El objetivo buscado al llevar a cabo este proceso no es otro que encontrar el mínimo número de dimensiones capaces de explicar de la mejor manera posible la información contenida en los datos.

[1] Diagrama obtenido de las diapositivas provistas para la asignatura SEMMA, ofrecida en el programa del Máster Minería de Datos e Inteligencia de Negocios

Una característica importante de este tipo de análisis es que todas las variables que intervienen cumplen el mismo papel, en el sentido de que son independientes, al no existir a priori una dependencia conceptual entre ellas.

Esta técnica consta de cuatros fases diferenciadas entre sí:

- Cálculo de la matriz en la que se expresará la variabilidad conjunta de todas las variables.
- Extracción del número óptimo de factores.
- Aplicación de una función de rotación a la solución obtenida, con el objetivo de facilitar su interpretación.
- Cálculo de las puntuaciones factoriales de cada uno de los individuos en la nueva dimensión.

Dos condiciones son indispensables para que tenga sentido la aplicación del análisis factorial, a saber, **parsimonia** (los fenómenos deben poder explicarse con el menor número de elementos posible) e **interpretabilidad**.

Se puede distinguir entre **Análisis Factorial Exploratorio**, donde no se conocen los factores "a priori", que se calculan mediante el análisis factorial, y A**nálisis Factorial Confirmatorio** donde se propone "a priori" un modelo, según el cual hay unos factores que representan a las variables originales. Siempre hay más variables que factores y se debe someter el modelo a comprobación. En el presente estudio, nos ocupará el primero de los casos mencionados.

Formalmente, el objetivo es resumir la información contenida en una matriz de datos con m variables, $(X_1, X_2, \dots, X_m)$, todas ellas tipificadas.Así, partiendo de las ecuaciones

$$X_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1k}f_k + u_1$$

$$X_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2k}f_k + u_2$$

$$\text{----------}$$

$$X_m = a_{m1}f_1 + a_{m2}f_2 + \cdots + a_{mk}f_k + u_m$$

Donde $|A| = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mk} \end{pmatrix}$ es la matriz de puntuaciones de cada factor con respecto a las variables originales, $f_i \forall i = 1, \dots, m$ son los factores o variables no observables que se buscan y los u_i, términos de error independientes e idénticamente distribuidos.

Para poder aplicar el análisis factorial, se suponeque los factores comunes están a su vez estandarizados ($E(f_i) = 0$; $Var(f_i) = 1$), los factores específicos tiene media 0 y están incorrelados ($E(u_i) = 0$; $Cov(u_i, u_j) = 0$ si $i \neq j$; j, $i = 1, \dots, k$) y que ambos tipos de factores también están incorrelados entre sí ($Cov(f_i, u_j) = 0, \forall i = 1, \dots, k$; $\forall j = 1, \dots, m$)

2.3. Análisis Cluster

El análisis *clúster* (Kiers & Rasson, 2000) utiliza la información de una serie de variables para cada sujeto u objeto y, conforme a estas variables, mide la similitud entre ellos. Una vez medida la similitud se organizan en grupos homogéneos internamente y diferentes entre sí. La "nueva dimensión" lograda con el *clúster* se aprovecha después para facilitar la aproximación "segmentada" de un determinado análisis.
A la hora de aplicar esta técnica, conviene tener en cuenta que tiene propiedades inferenciales, y que, como tal, los resultados obtenidos para una muestra sólo sirven para su diseño.

El análisis *clúster* trata de resolver el siguiente problema: dado un conjunto de n elementos, caracterizados por la información de $n$variables,X_j, se busca poder clasificarlos de manera que los individuos pertenecientes a un grupo o *clúster* sean tan similares entre sí como sea posible, atendiendo también a que los distintos grupos deben poder diferenciarse al máximo, en la medida de lo posible.

Este proceso consta de tres fases, que se pueden resumir en:

- Establecer un criterio de similitud para poder determinar una matriz de parecidos que nos permita relacionar la semejanza de los individuos entre sí.
- Aplicación de un algoritmo de clasificación con el que determinar la estructura de agrupación de los individuos.
- Especificación de dicha estructura, generalmente mediante el uso de diagramas arbóreos o dendrogramas.

Una vez seleccionadas las variables a considerar, cada uno de los individuos sujetos al análisis vendrá representado por los valores que tomen estas variables en cada uno de ellos. Este es el punto de partida de la clasificación. Para clasificar adecuadamente los individuos se debe determinar lo similares o disimilares (divergentes) que son entre sí, en función de lo diferentes que resulten ser sus representaciones en el espacio de las variables.
Para medir esta similitud, existen distintos índices, todos ellos con propiedades y utilidades diferentes, cuyo uso habrá que considerar en cada caso. La mayor parte de estos índices serán o bien, indicadores basados en la distancia (considerando a los individuos como vectores en el espacio de las variables); indicadores basados en coeficientes de correlación; o bien basados en tablas de datos de posesión o no de una serie de atributos.

En nuestro caso, nos basaremos en el concepto de distancia a la hora de medir la similitud entre los distintos objetos. Daremos el nombre de distancia entre dos individuos i y j a la medida $d(i,j)$, que indicará el grado de semejanza o desemejanza entre mencionados los objetos mencionados, en relación a un cierto número de características. Este valor, siempre debe cumplir:

1) $d(i,j)>0$
2) $d(i,i)=0$
3) $d(i,j)= d(j,i)$

En general, nos referiremos a la distancia euclídea como unidad de medida, la cual, además de las premisas mencionadas, verifica

4) $d(i,j) < d(i,t) + d(t,j)$
5) $d(i,j) > 0 \ \forall i,j$

En el presente estudio, se aplicará, en primer lugar, un método de *clúster* jerárquico, basado en la media de las variables en uso, que nos ayudará a determinar el número de *clusters* o grupos a crear, para luego aplicar un método no jerárquico, en el cual, una vez conocido el número ideal de *clusters* a crear, se formarán grupos homogéneos sin establecer relaciones entre ellos.

Utilizando los resultados obtenidos de los procesos correspondientes al análisis factorial, el *clúster* nos servirá como herramienta a la hora de comprobar cómo se comporta un determinado hotel con respecto a su competencia, entendiendo como tal a los hoteles pertenecientes al mismo grupo de todos los creados a través de los procesos *clúster*.

En este punto, se llevarán a cabo distintas comparativas, de modo que, se crearán grupos homogéneos atendiendo a las siguientes propiedades de los hoteles:

- Valoraciones otorgadas por los usuarios de *Booking*.
- Aspectos relacionados con la ubicación del hotel.
- Momento temporal.
- Número de estrellas del hotel.

Así, se estudiará a estudiar el comportamiento de un hotel atendiendo a las distintas agrupaciones obtenidas mediante el *clúster*, teniendo en cuenta distintas combinaciones de los aspectos mencionados.

2.4. Regresión Lineal

Los modelos de regresión lineal son modelos matemáticos usados para aproximar la relación de dependencia entre una variable dependiente Y, y las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como

$$Y_t = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Donde

- Y_t es la variable dependiente
- X_i son las variables explicativas o independientes
- β_i son parámetros que miden la influencia que las variables explicativas tienen sobre el regrediendo, siendo β_0 la intersección o término constante, β_i los parámetros respectivos a cada variable independiente y $n$el número de parámetros independientes a tener en cuenta en la regresión.
- ε es una variable aleatoria, que normalmente se supone Normal$(0,\sigma^2)$, independiente de cada observación.

El objetivo de la regresión es escoger unos valores determinados para los parámetros β_i, de modo que la ecuación anterior quede completamente especificada, de modo que pueda quedar determinada la relación existente entre la variable dependiente y las explicativas.

Para poder hacer esta determinación, deben cumplirse ciertos supuestos:

- Relación lineal entre variables.
- Independencia entre los errores obtenidos en la medición de las variables explicativas.
- Varianza constante en los errores. (*Homocedasticidad*)
- Esperanza matemática igual a 0 en los errores.
- La suma de todos los errores debe ser idéntica al error total.

En nuestro caso, la variable dependiente es cuantitativa, (precio), mientras que las variables explicativas independientes o covariables son de ambos tipos, cualitativas y cuantitativas.

Alguna de las covariables cualitativas son dicotómicas, como por ejemplo, la que nos dice si un hotel es estacional o no (¿sube su precio los fines de semana?).
En cuanto a las variables cualitativas con más de dos categorías, habrá puntos del proceso en que reciban un tratamiento especial, de modo que, para su inclusión en el modelo, se lleve a cabo una transformación de la misma variable en varias covariables dicotómicas ficticias (o de diseño), llamadas *dummy* de modo que cada una de las categorías se tomaría como categoría de referencia (para ello servirá de gran ayuda el uso de la macro *nombresmodbien* (Portela, 2015). De esta forma, cada categoría de cada covariable cualitativa no cuantitativa entraría en el modelo de forma individual. En general, si una variable categórica tiene *n* niveles, contaremos con *n-1dummys.*Otra opción a la hora de tratar con este tipo de variables consistirá en utilizar otras equivalentes de naturaleza cuantitativa, como se detallará en el epígrafe correspondiente a descripción de las variables.

A menudo jugaremos con la combinación de algunas categorías de variables cualitativas, de forma que se originará un efecto potenciador mayor o menor que el obtenido al sumar los distintos efectos por separado. Estas interacciones entre variables cualitativas son efectos en la regresión relativos a cruces de sus categorías, y serán denotados por el operador producto *. También se harán pruebas que incluyan combinaciones de variables cualitativas y cuantitativas.

Otras de las variaciones incluidas serán transformaciones llevadas a cabo sobre variables continuas para conseguir linealidad, o sobre variables discretas para simplificar los modelos.

A la hora de comparar los distintos modelos, se pueden considerar distintas medidas de ajuste. Todas ellas utilizan como base la suma de cuadrado de los errores SSE, y el total de la suma de cuadrados corregidos SST, donde:

$$SSE{:=}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \; SST{:=}\sum_{i=1}^{n}(y_i-\bar{y}_i)^2$$

Siendo

- y_i:= datos originales

- $\hat{y}_i$:= predicciones
- $\bar{y}_i$:= predicciones corregidas
- n:= número de observaciones

Así, las medidas de ajuste que utilizaremos serán, principalmente,

- $R2$:= 1- $\left(\frac{SSE}{SST}\right)$
- $R2\ ajustado$:= 1- $\frac{(n-1)(1-\mathrm{R}^2)}{(n-p)}$
- BIC:= $nln\left(\frac{SSE}{n}\right) + 2(p+2)q - 2q^2$, donde $q := \frac{n\sigma^2}{SSE}$
- AIC:= $nln\left(\frac{SSE}{n}\right) + 2p$
- SBC:= $nln\left(\frac{SSE}{n}\right) + pln(n)$
- $ASE := \frac{SSE}{n}$

Excepto $R2$ y $R2Adj.$, todas estas medidas se considerarán mejores cuanto más pequeñas sean. En nuestro caso, nos fijaremos en el *ASE* o error cuadrado promedio.

Las principales ventajas de la regresión lineal son:

- El análisis de regresión es una herramienta muy flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser numéricas y categóricas.
- Permite hacer una predicción del comportamiento de alguna variable en un determinado punto o momento.

2.5. Redes Neuronales

Las Redes Neuronales (Caicedo Bravo & López Sotelo, 2009) constituyen una herramienta muy potente de análisis, modelización y predicción. Su filosofía general es obtener modelos coherentes con la realidad observada, de tal modo que sean los datos los que determinen el comportamiento de la red, a través de sus estructuras o de parámetros internos. Las redes neuronales constituyen un método no paramétrico de obtención de datos.

Una red neuronal es un sistema de procesamiento de la información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. Consisten en un conjunto de elementos simples de procesamiento llamados nodos o neuronas conectadas entre sí por conexiones que tienen un valor numérico modificable llamado peso.

La actividad que una unidad de procesamiento o neurona artificial realiza en un sistema de este tipo es simple. Normalmente, consiste en sumar los valores de las entradas (*inputs*) que recibe de otras unidades conectadas a ella, comparar esta cantidad con el valor umbral y, si lo iguala o supera, enviar activación o salida (output) a las unidades a las que esté conectada. Tanto las entradas que la unidad recibe como las salidas que envía dependen a su vez del peso o fuerza de las conexiones por las cuales se realizan dichas operaciones.

La conexión de una red neuronal viene dada por un grafo con conexiones entre los nodos. Podemos distinguir entre nodos de entrada, de salida y ocultos. Los nodos de computación serán los de salida y los ocultos.

Los nodos se organizan formando capas. Cada nodo recibe un conjunto de entradas multiplicadas por su interconexión (peso), que son sumados y operados por una función de transferencia o activación, f_i, (que en general será diferente según se trate de unidades de la capa de salida o de unidades pertenecientes a la capa oculta), antes de transmitirse a la siguiente capa o como salida de la red como señal de salida, $f_i(y_i)$.

Dentro de una red neuronal, los componentes más importantes son los nodos o elementos de procesado, y sus interconexiones.
La capa input se conecta a la capa oculta mediante la función de combinación representada por $\sum$., en la que los pesos w_{ij} hacen el papel de parámetro a estimar.

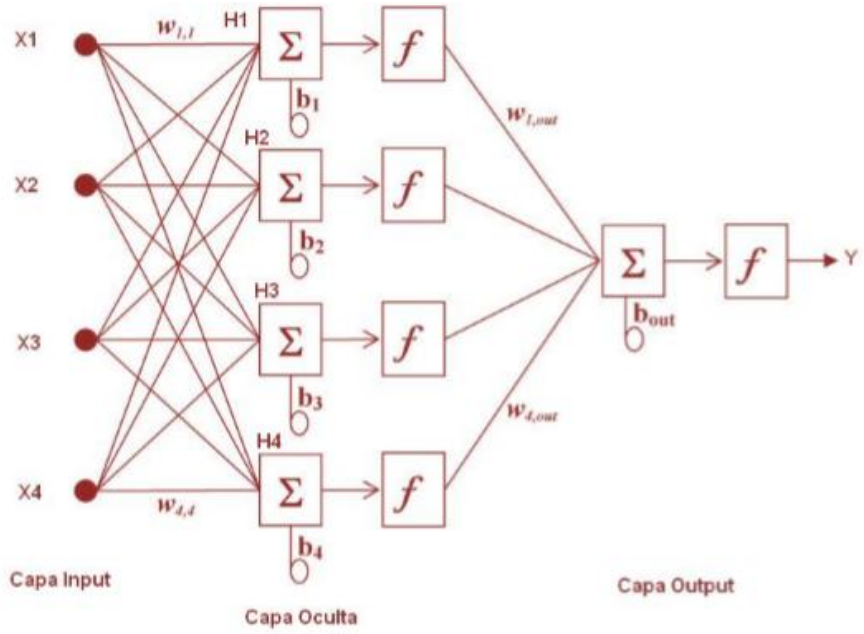

Figura 2. Ejemplo Red Neuronal

En este ejemplo, estamos ante una red neuronal con cuatro *inputs*, X_1, X_2, X_3, X_4; un *output* Y y una capa oculta con cuatro nodos ocultos H_1, H, H_3, H_4. Por otra parte, b_j se denomina como bias o sesgo.

La función de combinación más habitual es la lineal. Tras su aplicación, se ejecuta sobre cada nodo oculto la función de activación, representada por f (una de las más típicas es la tangente hiperbólica).

El objetivo computacional de las redes neuronales no es otro que estimar los valores de los parámetros w_{ij} y b_j.

Los métodos utilizados en las redes neuronales son técnicas de optimización numérica, que van variando los valores de los parámetros a estimar de manera iterativa hasta conseguir el objetivo de optimización, que suele ser la función de error en datos training o en datos de validación.

Las redes neuronales se basan en el Teorema de Aproximación Universal (Portela, 2015), cuya versión simplificada dice que si γ es una función no constante, acotada, monótona creciente y continua, entonces, dada cualquier función $f(x)$ en el hipercubo [0,1] y $\varepsilon > 0$, existen N y constantes $a_{i,} b_i$ y w_i en R^m tales que:

$$F(x) = \sum_{i=1}^{N} \alpha_i \gamma(w_i^T x + b_i)$$

$$|F(x) - f(x)| < \varepsilon$$

Volviendo a las redes neuronales, podemos interpretar este resultado como que, si existe una relación entre las variables *input* y la variable *output* y esta relación es no lineal y desconocida, es posible aproximar la función que define dicha relación, de modo que lo "único" que habría que decidir es el número de nodos, N, y la función de activación,γ. Esta elección se lleva a cabo mediante métodos iterativos, observando los valores que hacen óptimo el valor de la función objetivo sobre datos de validación.

Las Redes Neuronales presentan grandes ventajas que nos llevan a elegir este método de predicción en lugar de otros. Algunos de ellos son:

- Permiten solventar situaciones de no linealidad, o bien aquellas en las que la función entre variables input y output es desconocida.
- Facilitan el uso de datos de gran complejidad, como aquellos que presentan efectos temporales, o incluyen muchas variables categóricas o con datos censurados.
- Facilitan casos en los que el output es complejo, como por ejemplo, aquellos en que hay varias variables output simultáneas o de diferente tipo.
- Funcionan mejor para casos de clasificación que la regresión logística.
- No es necesario programar el aprendizaje, ya que las redes neuronales extraen sus propias reglas a partir de ejemplos reales, que quedan almacenadas y extendidas a lo largo de las conexiones.
- Son tolerantes al ruido, siendo capaces de abstraer las características esenciales de los datos y de generalizar estas correctamente.
- No son paramétricas ni necesitan hacer supuestos de la forma funcional de la función que van a aproximar.

Un detalle importante a tener en cuenta a la hora de utilizar este método predictivo es que, para poder usar las redes neuronales con garantías, se requiere que la fuente de datos tenga suficientes observaciones ya que, al no haber inferencia propiamente dicha, se necesitan datos test para validar el modelo.

En condiciones de linealidad en las relaciones, funciones de relación conocidas entre variables input y output, o bien cuando hay pocas observaciones, es recomendable abstenerse del uso de las redes neuronales. Además, no debemos olvidar las dos limitaciones que acompañan a este método:

- Es imposible determinar cómo se procesa internamente la información
- No existe una metodología clara y rigurosa a la hora de determinar el número de capas ocultas o de nodos que debe tener cada capa.

2.6. Árboles de Decisión

Un árbol de decisión (En línea 2016 b) es una forma gráfica y analítica de representar todos los eventos o sucesos que pueden surgir a partir de una decisión asumida en cierto momento. Ayudan a tomar la decisión "más acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones.

Los árboles de decisión constituyen una técnica de análisis discriminante no paramétrica que permite predecir la asignación de muestras a grupos predefinidos en función de una serie de variables predictoras.

Se trata de algoritmos iterativos que dividen los datos en regiones basadas en intervalos de las variables independientes. Así, los árboles tratan de hallar puntos de corte en las variables independientes que lleven a grupos de individuos con comportamiento homogéneo respecto a la variable respuesta y diferente entre los grupos. Por lo general, habrá que encontrar las regiones que minimicen una función de error dada. Como estamos ante un caso de regresión, en concreto será ASE (error cuadrado medio).

Un árbol se comporta como un análisis *clúster* (creación de grupos homogéneos y diferentes entre sí) y a la vez como un método predictivo. Por ello este tipo de técnicas cuando se orientan a crear grupos se denominan técnicas de segmentación.

El punto inicial (y único) de un árbol de decisión se denomina *nodo raíz*, y contiene el conjunto total sobre el que se va a proceder.
Un *nodo* es un subconjunto de variables, y puede ser terminal o no. Llamaremos *nodo terminal* o *padre*, a todo nodo que se divida en nodos descendientes.
Por el contrario, un nodo que no tenga descendientes, esto es, que no se divida en otro nivel, se llamara *nodo terminal*, y siempre tiene asignada una etiqueta de clase. Todas las observaciones a tratar van a parar a los nodos terminales. Cuando una observación de clase desconocida transita a través del árbol y va a parar a un nodo terminal, se le asigna la clase correspondiente a la etiqueta de clase adjunta a dicho nodo. Puede haber más de un nodo terminal con la misma etiqueta de clase.

En cada nodo, el algoritmo de generación del árbol tiene que decidir sobre qué variable es 'óptima' la partición a realizar. Necesitamos considerar cada posible división sobre todas las variables presentes en dicho nodo, enumerar después todas las posibles divisiones, evaluar cada una, y decidir cuál es la mejor siguiendo algún criterio.

A la hora de diseñar un árbol, hay que tener en cuenta muchos aspectos, prestando especial atención a los que se nombran a continuación:

- Establecer un buen criterio a la hora de elegir qué variable independiente o nodo va a ser la base en la siguiente división.
- Elección de los puntos de corte óptimos dentro de cada variable independiente o nodo.
- Elección de grupos de corte para variables independientes cualitativas con más de una categoría.
- Debe haber un número de observaciones mínimas para construir un nodo.

- Elección del criterio de parada-fin del algoritmo.
- Establecer un número máximo de nodos y divisiones.
- Establecer un criterio de tratamiento de *missings.*
- Establecer una decisión sobre el uso o no de datos de validación.

Frente a otros métodos, los árboles presentan una ventaja principal, que es su gran potencia descriptiva, la cual facilita mucho la interpretación de los resultados obtenidos. Además, hay que señalar que en ocasiones permiten descubrir interacciones y reglas que quedarían ocultas con el uso de otras técnicas, reglas que son candidatas a usar como variables *dummys* a la hora de aplicar otros métodos predictivos. Otras de las grandes ventajas que nos aportan los árboles de decisión es que las relaciones lineales no afectan demasiado a su comportamiento, como podrían hacerlo al de otros métodos predictivos, ya que, en los árboles, la única noción verdaderamente importante es el orden de las observaciones, además de que aportan medidas de importancia de las variables, y su propio proceso lleva incorporada una manera propia de tratar los datos *missings.*

No obstante, en contra de este método debemos asumir que se trata de una técnica compleja en su construcción y casuística, con poca eficacia predictiva, de poca fiabilidad y escasa generalización, ya que cada hoja o nodo del árbol es un parámetro, lo cual provoca modelos sobreajustados y algo inestables para la predicción.

En cualquier caso, estas desventajas son en ocasiones solventadas con técnicas de ensamblado de modelos, combinando muchos árboles, y con el apoyo de algunos de los algoritmos predictivos más potentes conocidos en la actualidad, que son *Bagging*, *Random Forest* y *Gradient Boosting.* Los dos últimos procesos mencionados, que serán usados en el estudio que nos ocupa, serán presentados a continuación.

Si bien en el presente estudio no se aplicará la técnica de árboles de decisión de forma directa, estará intrínsecamente usada en la aplicación de los algoritmos que se detallan en las siguientes páginas.

2.7. Random Forest

Random Forest (Breiman, 2001) es un algoritmo de combinación de árboles predictores, tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación del *Bagging*(En línea 2016 c)que consiste en incorporar aleatoriedad en las variables utilizadas para segmentar cada nodo del árbol.

A continuación se presenta el esquema de algoritmo *Random Forest.*

Dados los datos de tamaño N,

1) Repetir m veces i), ii), iii):
   (i) Seleccionar N observaciones con reemplazamiento de los datos originales
   (ii) Aplicar un árbol de la siguiente manera: En cada nodo, seleccionar p variables de las k originales, y de las p elegidas, escoger la mejor variable para la partición del nodo.
   (iii) Obtener predicciones para todas las observaciones originales N
2) Promediar las m predicciones obtenidas en el apartado 1)

Este algoritmo incorpora dos fuentes de variabilidad, de modo se mejora la capacidad de generalización. Además, esto reduce el sobreajuste, conservando en cualquier caso la facultad de ajustar bien las relaciones particulares de los datos.

En general, los parámetros a tener en cuenta en este algoritmo son los siguientes:

- El tamaño de las muestras, N, y el uso o no de reemplazamiento.
- El número de iteraciones a promediar, m.
- Características del árbol: número de hojas, profundidad, el número de divisiones máximas en cada nodo, el p-valor para las divisiones en cada nodo y el número de observaciones mínimas en cada rama-nodo.
- Número de variables a muestrear en cada nodo, p (si este número es igual a la cantidad inicial de variables, nos encontraríamos ante un caso de *Bagging*).

2.8. Gradient Boosting

El algoritmo *Gradient Boosting* (Friedman, 2001) consiste en repetir la construcción de árboles de regresión/clasificación, modificando ligeramente las predicciones iniciales cada vez, intentando ir minimizando los residuos en la dirección de decrecimiento, dada por el negativo del gradiente, de la función de error..

A medida que se van planteando diferentes árboles, el proceso ajusta las predicciones a los datos cada vez más, mejorando así la situación en que se plantea un único árbol.

Aunque no siempre es necesario, a menudo este proceso ha de ser monotorizado mediante *early stopping*[2], que nos ayudará a determinar el número de iteraciones, lo cual hace necesaria la existencia de datos de validación.

Aunque el esquema general del algoritmo es el mismo para casos de regresión o clasificación, surgen diferencias entre ambas situaciones a la hora de escoger la función base y la función de error.

[2]*Early Stopping:* es una forma de regularizar usada para evitar el sobrejuste esperado en procesos iterativos tales como descenso de gradiente
.

Esquema *Gradient Boosting*

1) Inicialmente, $f_0(x_0) = \ argmin \sum_{i=1}^{N} L(y_i, \gamma)$
2) $For\ m = 1\ to\ M$
   a. $For\ i = 1,2 \ldots N$:

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

   b. Crear árboles de regresión, recalcular los residuos r_{im}, a partir de las regiones temporales $R_{jm}, j = 1,2, \ldots J_m$.
   c. $For\ j = 1, 2, \ldots, J_m$,

$$\gamma_{jm} = \ argmin_{\gamma} \sum_{X_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

   d. Actualizar $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3) Finalmente, $\hat{f}(x) = f_M(x)$

En el caso de regresión, la función de error puede ser SCE[3] o ASE, mientras que para el caso de la clasificación, la función base establecida es $logit$, y $Deviance$ la función de error.
Se presenta el esquema para regresión, que será el utilizado en el presente estudio:

***Gradient Boosting* Regresión**

1) $\hat{y}_i^{(0} = \bar{y}$
2) Calcular el residuo actual, $r_i^{(m} = y_i - \hat{y}_i^{(m}$ (con la función de error dada, SCE, el residuo es igual al gradiente).
3) Actualizar $\hat{y}_i$ mediante $\hat{y}_i^{(m+1} = \hat{y}_i^{(m} + \ \alpha \hat{r}_i^{(m}$
4) Hacer $m \longrightarrow m + 1$. Volver al paso 1).

Es importante determinar las características propias de los árboles, el parámetro de regularización, α, y el número de iteraciones.

La gran ventaja ofrecida por este algoritmo es que es invariante frente a transformaciones monótonas, además de que lleva a cabo un tratamiento muy aceptable de datos *missings*, a lo cual hay que sumar su gran eficacia predictiva y la robustez que presenta frente a variables insignificantes, a colinealidad e interacciones ocultas. Además, este algoritmo es relativamente fácil de implementar. No obstante, hay que tener en cuenta que, como ocurre en todos los métodos basados en árboles, dependiendo

[3] SCE := $\frac{1}{2}\sum \left| y_i - f(x_i) \right|^2$

de la naturaleza y cantidad de datos a tratar, en ocasiones será preferible el uso de métodos más sencillos.

2.9. Comparación de Modelos

Una vez aplicadas las distintas técnicas mencionadas anteriormente y obtenidos los mejores modelos para cada una de ellas, el siguiente objetivo es hacer una comparativa final entre estos modelos con la finalidad de obtener el óptimo, considerando las ventajas e inconvenientes de cada modelo y el método empleado para su obtención.

Para comparar los modelos, obtenidos con la aplicación de las técnicas mencionadas, nos basaremos, en general, en el valor del error cuadrático medio, calculado en cada caso sobre los datos de validación. A la hora de llevar a cabo este tipo de comparativas en las que se ponen en tela de juicio modelos obtenidos mediante técnicas muy diferentes en las que la complejidad varía mucho, es importante contar con el matiz personal añadido por el investigador, que, más allá de buscar minimizar el error, ha de considerar hasta qué punto compensa escoger estructuras complejas para buscar esta disminución.

2.10. Software Empleado

Todos los métodos y procesos descritos en las páginas previas, se han llevado a cabo utilizando como soporte dos potentes módulos estadísticos proporcionados por SAS, como son:

- **SAS Enterprise Miner:** esta herramienta se tomará como apoyo fundamentalmente en las tareas propias del *Data Minning*, facilitándonos las primeras inmersiones y manipulación de los datos a tratar.
- **SAS Base*:*** mediante el código proporcionado en los Anexos, esta herramienta nos permitirá llevar a cabo los distintos procesos que abarcarán este estudio, proporcionándonos una fuente visual muy útil que facilitará la comprensión de los resultados obtenidos.

El principal motivo de elección del Software mencionado, es que ambos módulos han sido explotados en profundidad en el transcurso del Máster a que corresponde el estudio que nos ocupa.

3. Descripción de las Variables

Antes de empezar con el estudio Estadístico propiamente dicho, es de vital importancia conocer los datos que alimentarán la base con que trabajaremos, así como las variables en que están organizados. Por ello, a continuación se dará una breve explicación sobre cada una de las variables que formarán parte de la base de datos, detallando su naturaleza, características y el motivo de su integración en la fuente de datos, haciendo una distinción entre variables cualitativas y variables cuantitativas.

Hay que tener en cuenta que en nuestra base de datos contamos con variables estáticas, que son fijas o inherentes a cada hotel como por ejemplo su ubicación o cantidad de estrellas, y otras dinámicas, como el precio establecido por noche.

3.1. Agrupación por tipología de las variables

3.1.1. Variables Cualitativas

Las variables cualitativas se refieren a características o cualidades que no pueden ser medidas con números. Este tipo de variable permite distribuir los datos de acuerdo a características.
Dentro del conjunto de variables cualitativas, podemos distinguir entre variables nominales o categóricas y ordinales. Las variables nominales adoptan valores que no se pueden ordenar, caracterizando los datos por categorías de eventos mutuamente excluyentes y colectivamente exhaustivos. Por otro lado, las variables ordinales caracterizan, como su propio nombre indica, una relación de orden dentro de las categorías, atendiendo a una escala establecida.

Prestaremos especial atención a las variables dicotómicas 0-1 (caso especial de variable cualitativa nominal), que nos dirán si en una observación se cumple la característica representada por la variable en cuestión o no.

En las hojas correspondientes a Anexos Descriptivos se puede encontrar una tabla en la que se recogen todas las variables cualitativas que conforman nuestra base de datos y su número de categorías, además de otros datos de interés sobre cada una de ellas.

Las más importantes para el estudio serán, principalmente:

- ***Nombre Hotel***, que, como su propio nombre indica, no es más que la caracterización de cada hotel (por comodidad, a la hora de llevar a cabo los distintos procesos del estudio, y puesto que originalmente la base tiene los hoteles ordenados por su distancia al centro de la ciudad, nos referiremos a los hoteles con números, por ejemplo, llamaremos al *Hotel Europa*, primero de la lista, como 1. Esta información será recogida en la variable ***H***, también categórica, la cual, nos identificará los hoteles numéricamente, con el objeto de facilitar su tratamiento.
- Número de **ESTRELLAS** de cada hotel.
- **ZONA** en que se encuentra cada establecimiento.

- **FECHA**¸ indica con valores numéricos, que comprenden del 1 al 31, y representan el día del mes de Agosto de 2016 a que corresponde cada observación.

Otras variables cualitativas que conforman nuestra base de datos son:

- **FECHAINSCRIPCION**, variable que indica el día en que un determinado hotel se registró en el sitio web *Booking.*
- **CODPOSTAL**, que indica el código postal correspondiente a cada establecimiento.
- **DIA**: esta variable indica a qué día de la semana corresponde cada observación. Puesto que esta variable está fuertemente relacionada con FECHA, se estudiará hasta qué punto es necesario contar con ambas, ya que cabría pensar que, a la hora de un hotel fijar el precio de sus habitaciones, no se fija tanto en el día del mes, sino en el día de la semana en cuestión. Con esto, se busca explicar, por ejemplo, si todos los sábados del mes tienen un comportamiento parecido, o si, por el contrario, el primer sábado del mes difiere en gran medida del último.
- **FINDE**: esta variable toma tres valores, que se detallan a continuación:
  - 0, si el día siguiente es laboral.
  - 1, si el día en cuestión es laboral, y el día siguiente no.
  - 2, si el día en cuestión y el siguiente no son laborales.

  De esta forma, no sólo se contemplan los fines de semana, sino también los posibles festivos de cada mes.

Por otra parte, la base de datos que nos ocupa incluye distintas variables dicotómicas de gran importancia, como lo son:

- **ESTACIONAL***:* un hotel será estacional si podemos afirmar que su precio aumenta todos los fines de semana con respecto a los días de diario.
- **OFERTA:** $\forall i, j, 1 \leq i \leq 31, 1 \leq j \leq 278$, el valor j-ésimo de una variable o_i nos indicaría si el precio publicado por el hotel j para la noche i es una cantidad estándar prefijada por el establecimiento o si, por el contrario, está sujeto a alguna oferta especial. Así, nuestra variable OFERTA aglutinará esta información para todos los hoteles y noches que nos ocupa, tomando los valores 1, si el precio se corresponde con una oferta, y 0 en caso contrario.

Pese a la naturaleza inicial de las variables mencionadas, en general, en algunos de los procesos de predicción que nos ocuparán, se llevarán a cabo distintas pruebas en que estas variables serán tratadas como cuantitativas, con el objetivo de simplificar su integración y tratamiento en los distintos modelos. Para ello, a cada valor tomado por cada una de las variables categóricas, se le asignará el promedio de la variable a estimar, el *PRECIO*, para cada uno de los niveles, calculado sobre el conjunto de datos de entrenamiento. Por ejemplo, el valor *Chamberí* de la variable *ZONA*, sería sustituido por el promedio del precio de los hoteles de esta zona, en los días del 1 al 21 de agosto. Esta "variación" de la variable inicial será nombrada con el nombre de la misma, seguida del sufijo *–prom* (indicando que se trata de un promedio del precio).

3.1.2. Variables Cuantitativas

Las variables cuantitativas representan características que pueden ser expresadas mediante números, a las que se le pueden aplicar operaciones aritméticas. Dentro de este grupo, podemos distinguir dos tipos: variable cuantitativa discreta, es aquella que sólo toma un número finito de valores entre dos valores cualesquiera de una característica, y variable cuantitativa continua, que es aquella que puede tomar un número infinito de valores entre dos valores cualesquiera de una característica.
Las variables cuantitativas que nos ocupan son las siguientes:

- $\boldsymbol{p_i}$, $\forall i,j, 1 \leq i \leq 31, 1 \leq j \leq 278$, el valor j-ésimo de una variable p_i se corresponde con el precio publicado por el hotel j-ésimo para la noche$i$. Todas estas variables son de naturaleza continua, y, para su uso en el estudio que nos ocupa, todas ellas serán organizadas y recogías en la base de datos por la variable **PRECIO.**
- **PROMEDIO:** esta variable nos indica el precio medio de cada hotel, y también es continua.
- **COMENTARIOS**: esta variable cuantitativa discreta indica la cantidad de comentarios recibidos por cada hotel en el sitio web *Booking.* Junto con la variable *FECHAINSC,* nos permite medir la "popularidad" de cada establecimiento.
- **Distintas Valoraciones,** otorgadas por los usuarios de *Booking* a los distintos hoteles, calificando los siguientes varios aspectos de cada establecimiento, como lo son:
    - **Relación CALIDADPRECIO.**
    - **CONFORT**.
    - **INSTALACIONES y SERVICIOS**.
    - **LIMPIEZA.**
    - **PERSONAL.**
    - **UBICACIÓN.**
    - **WIFI**.
    - **Valoración Global.**

  Las valoraciones de los distintos aspectos mencionados, se recogerán en sendas variables continuas, una por cada matiz nombrado.

3.2. Agrupación por periodicidad de las variables

3.2.1. Inherentes a cada Hotel (Estáticas)

Estas variables nos dan información invariable en el tiempo sobre cada uno de los hoteles. Estas variables nos describen los hoteles. Si bien en la realidad las valoraciones recogidas en *Bookinkg* varían de acuerdo a la intervención de los usuarios, como los datos están recogidos en un instante temporal fijo, estas variables serán consideradas estáticas en el presente estudio.

- HOTEL
- ESTRELLAS
- FECHAINS

- ZONA
- COMENTARIOS
- USABLE
- ESTACIONAL
- CALIDADPRECIO
- CONFORT
- INSTALACIONESSERVICIOS
- LIMPIEZA
- PERSONAL
- UBICACIÓN
- VALGLOBAL

3.2.2. Inherentes al dúo Fecha/Hotel (Dinámicas)

Este grupo de variables nos permiten ver las variaciones de precios atendiendo al día o la oferta a que esté sujeto cada hotel en un momento determinado.

- OFERTA
- FECHA
- DIA
- FINDE
- PRECIO

3.3. Primer Acercamiento a los Datos.

Antes de adentrarnos en los objetivos principales del estudio que nos ocupa, es importante conocer los datos en uso, así como la forma en que están relacionados. En este punto, intentamos encontrar argumentos estadísticos que nos permitan afirmar relaciones, que en un principio, desde un punto lógico y con un conocimiento mínimo del campo que se está explorando, nos parecen evidentes, por ejemplo *en general, un hotel de 4 estrellas es más caro que uno de 2*, o bien *en general, los hoteles del centro encarecen su precio los fines de semana*.

Así, en este punto, buscaremos ilustrar estadísticamente algunas de las relaciones principales existentes entre las variables más significativas. Para ello, aplicaremos modelos básicos de regresión (en los siguientes apartados se hará un estudio de regresión en profundidad), para casos en que busquemos relaciones entre variables continuas, y diseño de experimentos si la variable a relacionar con el precio fuera categórica. En ambos casos, nos apoyaremos en el proceso de SAS *proc GLMSelect.*

En los Anexos correspondientes a Análisis Descriptivos, podemos encontrar más datos acerca de lo que se contará en este apartado, como pueda ser la matriz de correlación de las distintas variables con respecto al Precio.

Por ejemplo, una relación que parecería bastante clara, es la existente entre el precio de un hotel y sus estrellas. Si calculamos el coeficiente de correlación entre estas variables, obtenemos 0.49625 y un $p - valor$ inferior a 0.001, lo cual nos permite afirmar que esta relación es lo suficientemente significativa. Además, el siguiente gráfico (Figura 3) nos muestra cómo se comporta el precio atendiendo al número de estrellas, y en la Tabla 11, *Comparisons significant at the 0.05 level are indicated by* ***, relativa a la variable *ESTRELLAS*, podemos ver que cada categoría de esta variable presenta un comportamiento con respecto al precio totalmente independiente del resto.

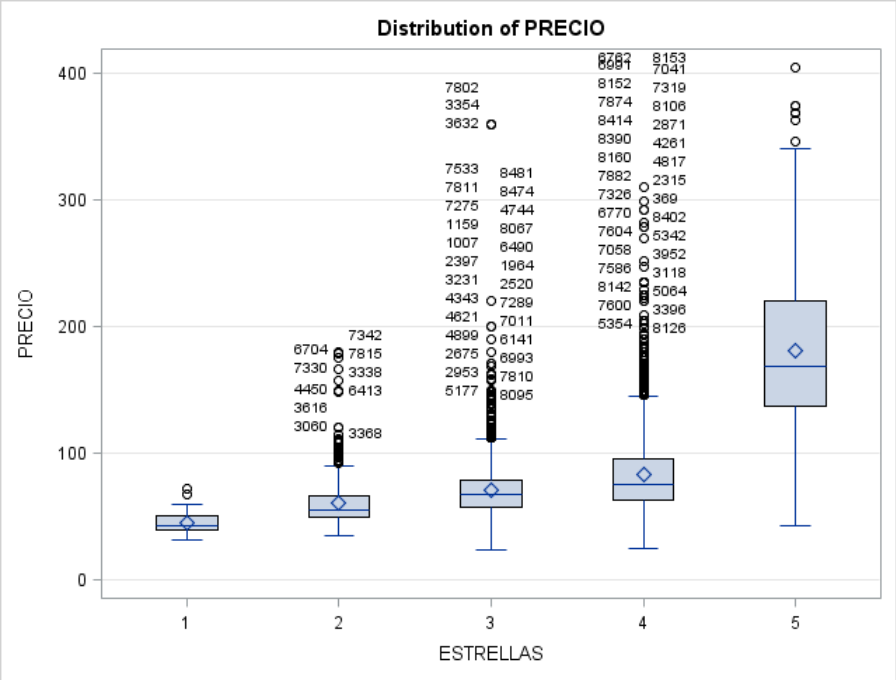

Figura 3. Distribución del PRECIO atendiendo a la variable ESTRELLAS

Otro aspecto que podría resultar de gran utilidad conocer, es, hasta qué punto las calificaciones con que los usuarios de *Booking* califican a los distintos hoteles, afectan a la hora de que estos fijen sus precios. En este caso, nos enfrentamos a un caso de relación entre variables continuas, que mediremos mediante el coeficiente de correlación de Pearson entre cada una de ellas, que queda reflejado en la matriz de correlaciones que se puede encontrar en los Anexos (Tabla 12). En ella podemos ver que el coeficiente de correlación del precio con todas y cada una de las variables de puntuación es mayor que 0.4 y el $p - valor$ es inferior a 0.001, lo cual, nuevamente, nos permite afirmar que existe una relación importante entre el precio de un hotel y la opinión de sus clientes.

Por otra parte, parece natural preguntarse la relevancia del día de la semana a la hora de establecer un precio. En la Tabla13, *Comparisons significant at the 0.05 level are indicated by* ***, correspondiente al estudio de la variable *DIA*, se puede observar que, al 95% de confianza, las únicas categorías de la variable *DIA* que son comparables entre sí, se corresponden con los días de domingo a jueves, teniendo los viernes y sábados, comportamientos diferentes. Reforzando este argumento, podemos ver cómo los siguientes gráficos de cajas y bigotes nos muestran los "picos" que aparecen con la llegada de los fines de semana.

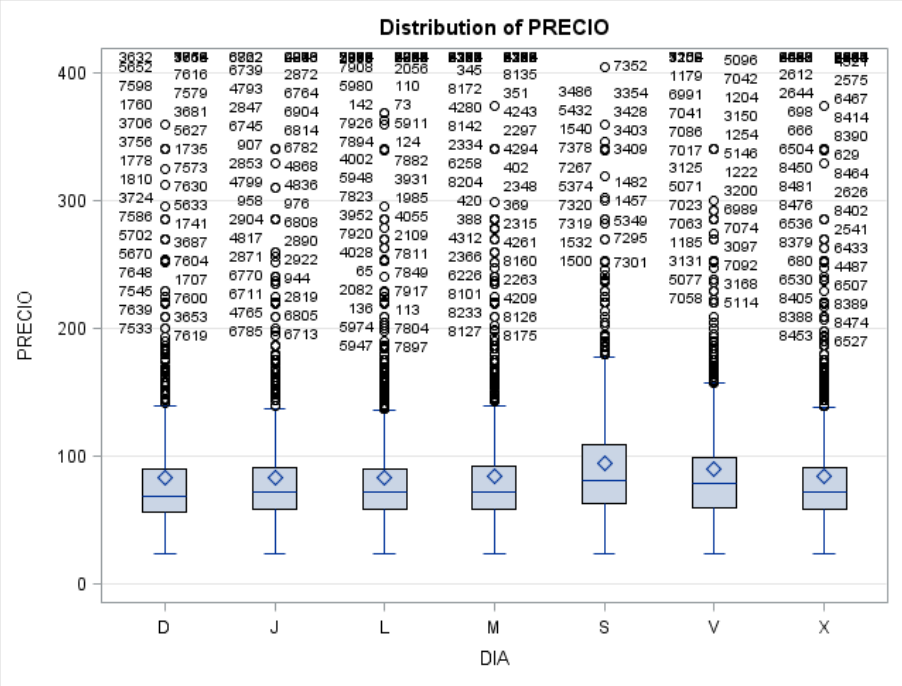

Figura 4. Distribución del PRECIO atendiendo a la variable DIA

Afinando más en las posibles cuestiones que podemos plantearnos, podría surgir la duda de hasta qué punto es necesario conocer el día del mes correspondiente, si ya sabemos el que el día de la semana tiene una importancia relevante a la hora de explicar el precio. Para responder a esta cuestión, se estudia la importancia de la variable *FECHA* a la hora de predecir todo aquello que no queda explicado con *DIA*, esto es, el residuo correspondiente. El resultado obtenido nos dice que, dado que el *p-valor* es 0.0012, bajo un nivel de significaciónα =0,05, los datos no muestran evidencia estadísticas para rechazar la hipótesis nula, que dice que el parámetro de la variable *FECHA* en el modelo no tiene importancia explicativa en presencia de la variable *DIA*, esto es, que su comportamiento no afecta a la hora de explicar el precio. Como conclusión a esto, no podemos garantizar que la variable *FECHA* explique nada relevante en presencia de la variable *DIA*. No obstante, se harán distintas pruebas, variando el uso de una u otra variable.

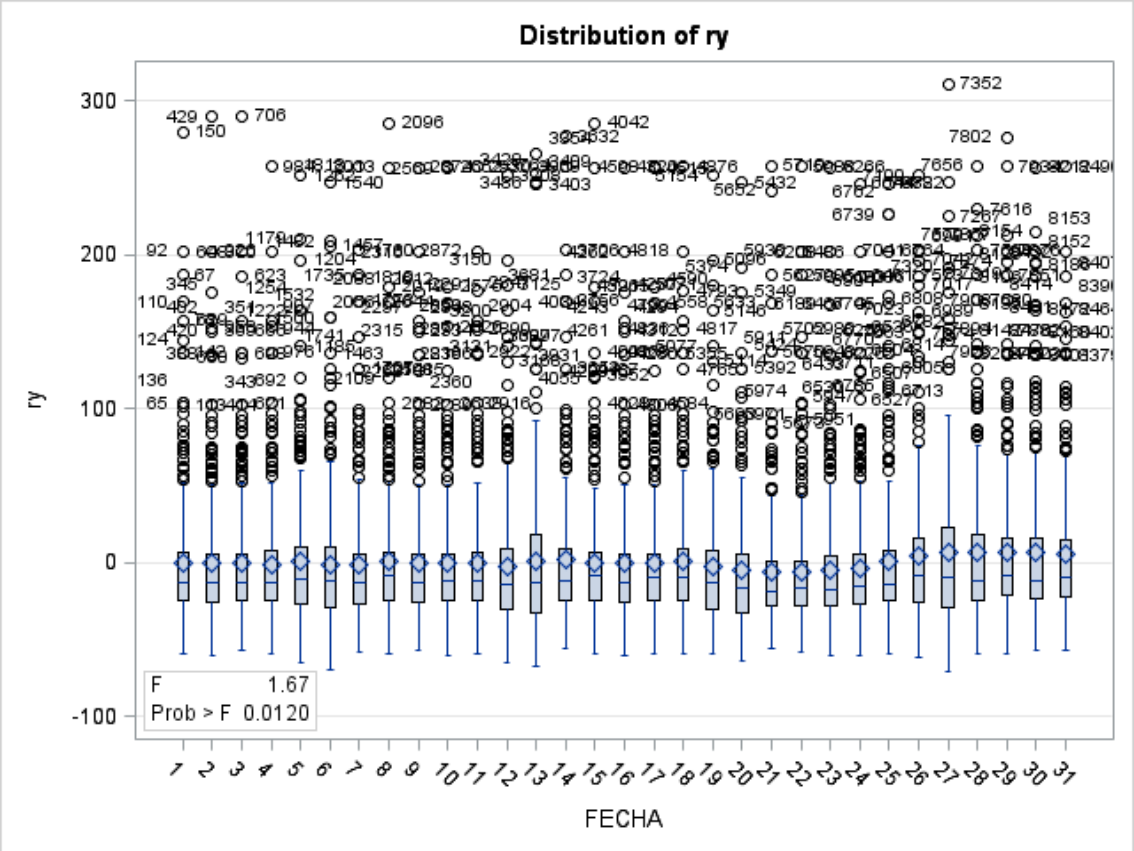

Figura 5. Distribución con respecto a la variable FECHA de los residuos del PRECIO no explicados por DIA

Por otra parte, parece razonable preguntarse por la influencia de la zona sobre el precio. Como se puede observar en la Figura 6, las zonas en las que el precio se dispara son Retiro y Salamanca, lo cual, ante un mínimo conocimiento de la ciudad de Madrid, debe resultar bastante evidente (en este punto, aunque pueda llamar la atención el comportamiento de los hoteles situados en Villa de Vallecas, por su alto precio, es importante tener en cuenta que en esta zona sólo hay dos hoteles, siendo uno de ellos de 4 estrellas, el cual provoca la subida mostrada en el gráfico).

Bajo un nivel de significación de α =0,05, podemos afirmar que algunas de las zonas "parecidas" o "comparables" entre sí son Chamberí y el Centro, o Chamartín y Barajas. En los Anexos correspondientes a este epígrafe se puede encontrar la tabla de comparaciones significativas con respecto a los distintos valores que toma la variable *ZONA*.

Profundizando más en este punto, al igual que previamente se plantea la cuestión de la relevancia de la variable *FECHA* en presencia dela variable *DIA*, en este contexto cabría preguntarse hasta qué punto el código postal nos aporta información que se pueda escapar en caso de contar únicamente con la variable *ZONA* (los distintos códigos postales de la ciudad se organizan en forma de distritos).

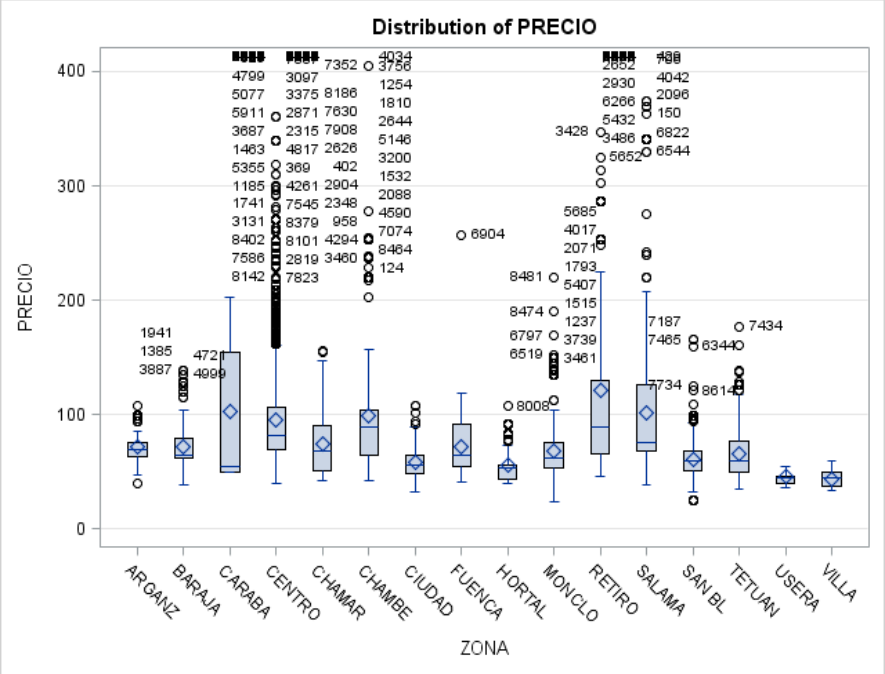

Figura 6. Distribución de los residuos con respecto a la variable ZONA

El resultado obtenido nos dice que, dado que el *p-valor* es menor que 0.001, bajo un nivel de significación α =0.05, los datos no muestran evidencias estadísticas que nos permitan rechazar la hipótesis nula, que en este caso dice que la variable *CODPOTSAL* no tiene importancia explicativa en presencia de la variable *ZONA*. En otras palabras, no podemos garantizar que la variable *CODPOSTAL* explique algo que se escape del dominio de la variable *ZONA*¸ esto es, que afecten a la hora de analizar los residuos correspondientes. En cualquier caso, las diferentes pruebas se llevarán a cabo considerando una u otra variable.

Además de todas las relaciones mencionadas, previamente al estudio de predicción, se lleva a cabo un análisis profundo de todas las variables iniciales, y de su nivel de significación a la hora de explicar la variable dependiente en cuestión, con el objetivo de evitar pruebas que pudieran incluir variables "poco importantes".
Este análisis prácticamente permitiría descartar directamente las variables *FECHAINSC* y *COMENTARIOS*, puesto que no aportan información relevante sobre la variable dependiente que nos ocupa. En cualquier caso, se llevará a cabo la obtención de distintos modelos que incluyan estas variables, y se comprobará que, en esencia, no mejorarán a otras estructuras que no las contengan.

4. Modelos de Predicción

4.1. Fragmentación de la Información

En este tipo de procesos analíticos hay que tener en cuenta que tan importante es el desarrollo de la técnica de predicción, como su posterior validación y correspondientes pruebas. Para ello, se procede a dividir la base de datos original en tres archivos distintos, cada uno de los cuales contará con la información relativa a cada uno de los pasos mencionados. Así, la división quedaría como sigue:

- **Aprendizaje** (datos *training*)*:* estos datos, comprenderán los precios, ofertas y demás aspectos de los distintos hoteles, de los días 1 al 21, ambos inclusive, y serán utilizados por cada uno de los modelos de predicción que se van a llevar a cabo para predecir los valores de los parámetros inherentes a cada modelo.

- **Validación** (datos *validation*): esta información será utilizada para decidir cuál de todos los modelos de predicción llevados a cabo comete un error menor en la estimación, siendo así el mejor. Estos datos estarán comprendidos por la información relativa a los días 22 al 28.

- **Prueba** (datos *test*): conjunto de datos formado por la información de los días 29, 30 y 31 de Agosto, y que será utilizado para predecir de forma no sesgada el grado de acierto de los modelos seleccionados como mejores en validación.

Además de utilizar esta fragmentación de la información para garantizar el correcto comportamiento de los distintos modelos de predicción obtenidos, se aplicará la técnica de validación cruzada con el objetivo de tener otra vía de garantizar que los resultados obtenidos son independientes a los datos utilizados para la construcción de las distintas estructuras predictivas.

4.2. Análisis Factorial

Un paso previo a llevar a cabo antes de profundizar en el estudio predictivo consiste en la realización de un análisis factorial, mediante el cual se buscará explicar posibles correlaciones entre las distintas variables. Este proceso nos permitirá establecer una relación entre variables y atributos que a primera vista no son aparentemente relacionados, proporcionando así una estructura interna y las correlaciones que subyacen de ella.

Para llevar a cabo este tipo de análisis, nos apoyaremos en los resultados obtenidos con la aplicación del procedimiento de SAS *proc factor.* Para ello, se le asigna una parrilla de distintos valores a los parámetros que forman parte de este proceso, tal y como se detalla a continuación.

- **Conjunto de Variables**: a la hora de llevar a cabo este tipo de análisis, es importante tener en cuenta que las variables en uso deben ser de naturaleza continua. Por ello, como paso previo, hay que "traducir" las variables cualitativas implicadas, de modo que para cada una de ellas se creará una variable continua que tomará como valor, el promedio del precio con respecto a cada uno de los niveles de la variable cualitativa a la que represente, acción detallada en el epígrafe correspondiente a descripción de variables.
- **Priors**: este parámetro selecciona el método utilizado para el cálculo de las estimaciones previas. Se llevarán a cabo pruebas utilizando las siguientes variantes para este parámetro:
    - *SMC*: este método establece como estimación previa para cada variable el cuadrado de su correlación múltiple con respecto al resto de variables.
    - *MAX*: este método establece como estimación previa para cada variable el máximo de todas sus correlaciones con el resto de variables.
    - *RANDOM:* este método establece como estimaciones previas un conjunto aleatorio de números distribuidos uniformemente entre 0 y 1.
- **NFactors:** este parámetro representa el número máximo de factores a crear. Como es de esperar, esta cantidad no puede ser mayor que el número de variables implicadas. Para este parámetro no se establece un conjunto de valores fijos con que hacer pruebas, sino que a medida que se hacen pruebas, se ajusta con respecto a los resultados obtenidos.
- **Rotate**: este parámetro representa a la función o método de rotación, que nos permitirá transformar la matriz de correlaciones inicial sin alterar sus propiedades matemáticas. Se llevarán a cabo pruebas con las siguientes funciones de rotación:
    - *VARIMAX*: se trata de un método de rotación ortogonal que minimiza el número de variables que tienen saturaciones altas en cada factor. Simplifica la interpretación de los factores.
    - *QUARTIMAX:* este método de rotación minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.
    - *PARSIMAX*: este método lleva a cabo una rotación ortogonal, en el que se minimizar el número de factores, además de simplificar su interpretación.

  Todos los métodos utilizados para llevar a cabo la rotación son de naturaleza ortogonal, puesto que en los primeros pasos de este proceso, se observa que la correlación entre los distintos factores es nula.

Tras llevar a cabo distintas pruebas, resultantes de asignar una parrilla de distintos valores a los distintos parámetros que intervienen en este proceso, se establece, que, seleccionando *SMC* como método para el cálculo de las estimaciones previas y aplicando *varimax* como función de rotación, obtenemos tres factores, con los que "explicar" la información recogida en nuestras variables de partida de una forma bastante acertada. En la siguiente tabla se muestra cómo estos nuevos factores representan o explican a las variables iniciales.

VARIABLE	Factor1	Factor2	Factor3
CALIDADPRECIO	0.80870	-0.08909	-0.00370
CONFORT	0.95373	0.07295	0.00459
INSTALACIONESSERVICIOS	0.97349	0.10504	0.00089
LIMPIEZA	0.91507	0.17511	-0.00347
PERSONAL	0.77539	0.30400	-0.00266
UBICACION	0.26124	0.82697	-0.00163
VALGLOBAL	0.50616	0.26383	0.00622
WIFI	0.56756	0.27744	-0.00571
CPPROM	0.12304	0.79737	0.00119
ZONAPROM	0.10352	0.77536	-0.00214
DIAPROM	-0.00045	0.00627	0.97462
FECHAPROM	-0.00083	0.00777	0.71031
OFERTAPROM	0.00097	-0.09120	0.12951
FINDE	-0.00025	0.00501	0.97807
FECHAINSC	0.17525	-0.10394	0.00001
COMENTARIOS	0.02185	-0.04212	0.00152

Tabla 1. Factores

Aunque en principio no es trivial asignar una interpretación concreta a los factores obtenidos, observando las celdas sombreadas de color amarillo, se puede observar que, cada uno de nuestros tres factores parece explicar un conjunto de aspectos determinado:

- **Factor 1**: podemos interpretar este factor como un resumen de las valoraciones de los usuarios de *Booking*
- **Factor 2**: este factor podría interpretarse como un explicación de la ubicación de cada hotel.
- **Factor 3**: este factor recoge información inherente al momento temporal del alquiler a realizar.

Así, se puede asumir que nuestras variables quedan muy acertadamente explicadas con los tres factores presentados. La única característica relevante que podría echarse de menos en este momento, es el número de estrellas de cada hotel, pero este punto será solventado en los siguientes pasos del análisis.

4.3. Regresión

Puesto que nuestra variable a predecir, el precio, es cuantitativa, estamos ante un caso de Regresión Lineal. A continuación se detallarán los pasos a seguir, los modelos obtenidos y las conclusiones que se pueden extraer en este punto del estudio. En las hojas correspondientes a Anexos Analíticos, se pueden encontrar gráficos y tablas que facilitarán la comprensión del proceso.

4.3.1. Obtención de modelos. Búsqueda del modelo óptimo.

La búsqueda del mejor modelo de Regresión Lineal se lleva a cabo de forma empírica en un proceso prueba-error, consistente en la obtención y comparación de distintos modelos mediante la aplicación de distintos métodos de selección y ajuste de parámetros. Con ello, se han obtenido un total de 185 modelos distintos, todos ellos susceptibles de variaciones. Es importante recordar que esta búsqueda se llevará a cabo sobre el conjunto de datos de entrenamiento o aprendizaje, para después valorar su funcionamiento sobre los datos de validación.

El esquema general de pasos a seguir es el que se presenta a continuación:

- Variación del conjunto de **variables** a utilizar. Es importante jugar con la presencia de los factores obtenidos en el punto anterior, que ayudarán a evitar futuros problemas de multicolinealidad.
- Aplicación de distintos métodos de **selección de variables**, a saber, *Stepwise* (*SignificanceLevel*, AIC,SBC, SL), *Backward* (*SignificanceLevel*, AIC, SBC, SL) y *Forward* (AIC, SBC, SL). En la aplicación de los métodos mencionados, se asignarán distintos valores a los posibles criterios de entrada o salida de las variables en el modelo, que atenderán al nivel de significación de las mismas. Estos valores serán 0.1, 0.01, 0.005, 0.001, 0.0001.
- Variación de los parámetros de **aleatorización** ¿Aportan mejoras al modelo? ¿En qué medida dependen los modelos obtenidos de las observaciones utilizadas? En este punto se harán distintas pruebas con la semilla de aleatorización, y se buscarán los efectos que más se repitan, con el objetivo de obtener un modelo lo más estable posible.
- Profundización en el tratamiento de **variables categóricas**; creación de *dummys* para su incorporación en el modelo. Es importante evitar categorías poco representadas que pueden dar problemas en la construcción del modelo y la posterior predicción. Una buena idea podría ser agrupar dichas categorías o bien, sustituir las variables categóricas por sus equivalentes cuantitativas, mediante el proceso descrito en el epígrafe dedicado a la descripción de las variables.
- **Interacciones** entre variables. ¿Nos ayudan a obtener modelos mejorados?

- **Transformaciones** de las variables, con el objetivo de mejorar la linealidad entre ellas. Las transformaciones que se probarán son $\frac{1}{x+1}, \log(x+1), \sqrt{x}, x^2$ y x^3, siendo x la variable a transformar. ¿Optimizan el modelo estas transformaciones?
- Elección del **modelo óptimo**

Tras llevar a cabo este proceso de búsqueda y análisis, se concluye, con la ayuda de los datos de validación, que el modelo candidato a óptimo es el que contiene las siguientes variables:

FACTOR1 CALIDADPRECIO UBICACIÓN VALGLOBAL CPPROM FECHAPROM OFERTAPROM ESTRELLASPROM

Este modelo, con sólo 9 parámetros a estimar, se ha obtenido aplicando el método de selección de variables *Stepwise*, mediante la selección *SignificanceLeve.*, fijando como criterio de entrada y salida de variables *sle*=*sls*= 0.001.

Con un error cuadrático medio de 824.15 calculado sobre los datos de validación, el presente modelo difiere en menos de un 2% del mejor modelo obtenido en cuanto a efectividad se refiere.

Varias razones nos han motivado a seleccionar este como modelo óptimo, además del buen aspecto que presentan sus medidas de ajuste. Más allá de tener una efectividad totalmente aceptable, el modelo seleccionado como óptimo posee una característica principal que ha motivado la elección, y es su sencillez. A medida que se dan los pasos mencionados, se comprueba que algunas tareas, como la búsqueda de interacciones entre variables, traen consigo más dificultades que beneficios, ya que en general, no aportaban una mejora al modelo base, y sí que pueden venir acompañadas de cierta inestabilidad.
Así, la elección de este modelo nos evita procesos tan costosos a la hora de predecir como lo son la creación de *dummys* o la agrupación de categorías.

Con el proceso ensayo-error llevado a cabo, se observa que las técnicas descritas en el esquema anterior, aportan, en el mejor de los casos, mejoras ínfimas a la efectividad de los modelos, lo cual nos conduce a la decisión tomada. La combinación de todos los detalles mencionados, convierte a nuestro modelo en una estructura robusta y estable pese a la variabilidad de datos. Además, el pequeño número de parámetros a estimar, evitará que este modelo esté sobreparametrizado, matiz que nos permite observar que cada una de las variables que forman parte de él tiene un papel claro y directo, fácilmente interpretable.

Una vez seleccionado el modelo mencionado como mejor de entre todos los obtenidos con la combinación de parámetros seguida, se procede, mediante un acto de prueba-error, a hacer transformaciones elementales en las variables que lo componen, intentando mejorar en algo su dispersión. Nuevamente, se observa que este trámite es evitable, ya que en ninguno de los casos se obtienen mejoras significativas tras transformar alguna de las variables independientes. Si bien en algunos casos excepcionales la transformación de una variable mejora en cierto punto su dispersión, se concluye que esta leve aportación al modelo no compensa el proceso necesario para llegar a ello. En la parte relativa a regresión de los Anexos Analíticos se pueden

observar gráficas de la dispersión de las variables que constituyen el modelo seleccionado antes y después de ser sometidas a diversas transformaciones.

Así pues, tras numerosas pruebas y comparativas, se concluye que el modelo óptimo es el presentado anteriormente. Matemáticamente, este modelo quedar expresado por la siguiente ecuación**:**

$$\begin{aligned}\mathbf{PRECIO} = \boldsymbol{\mu_0} + \boldsymbol{\mu_1}\mathbf{FACTOR1} + \boldsymbol{\mu_2}\mathbf{CALIDADPRECIO} + \boldsymbol{\mu_3}\mathbf{UBICACION} \\ + \boldsymbol{\mu_4}\mathbf{VALGLOBAL} + \boldsymbol{\mu_5}\mathbf{CPPROM} + \boldsymbol{\mu_6}\mathbf{FECHAPROM} \\ + \boldsymbol{\mu_7}\mathbf{OFERTAPROM} + \boldsymbol{\mu_8}\mathbf{ESTRELLASPROM}\end{aligned}$$

Donde, $\forall \mathbf{i}, \boldsymbol{\mu_i}$son los parámetros estimados en el proceso de regresión. Tomando los valores obtenidos para los mismos en la obtención del modelo óptimo, esta ecuación tomaría la siguiente forma:

$$\begin{aligned}&\mathbf{PRECIO} = \\ &\mathbf{16.73} + \mathbf{20.2}\boldsymbol{FACTOR1} - \mathbf{28.37}\boldsymbol{CALIDADPRECIO} + \mathbf{13.23}\boldsymbol{UBICACION} - \\ &\mathbf{8.35}\boldsymbol{VALGLOBAL} + \mathbf{0.27}\boldsymbol{CPPROM} + \mathbf{0.75}\boldsymbol{FECHAPROM} + \\ &\mathbf{1.12}\boldsymbol{OFERTAPROM} + \mathbf{0.76}\boldsymbol{ESTRELLASPROM}\end{aligned}$$

4.3.2. Interpretación del modelo óptimo.

A continuación, analizaremos el modelo seleccionado como óptimo, estudiando el comportamiento de sus variables con respecto a la variable a predecir, el precio.

Para empezar, podemos agrupar estas variables atendiendo a su significado:

- Relativas a las valoraciones de los clientes, contamos con *FACTOR1*, *CALIDADPRECIO*, *UBICACIÓN* y *VALGLOBA*
- Relativas al punto en que esté situado el hotel, contamos con *CPPROM.*
- Relativa al momento temporal, contamos con *FECHAPROM.*
- Relativa a características estáticas de los hoteles, tenemos *ESTRELLASPROM.*
- Relativa a la fluctuación del comportamiento de los hoteles, contamos con la variable *OFERTAPROM.*

Incluso, podríamos plantearnos tratar la variable *UBICACIÓN* como una caracterización de la zona en que se encuentra cada hotel, más que como una valoración cualquiera.
En cualquier caso, se observa que nuestro modelo recoge prácticamente todas las características por las que podríamos "definir" un hotel.

En las hojas correspondientes a los Anexos Analíticos de Regresión, se incluyen imágenes que muestran, de forma gráfica, el comportamiento de la variable *PRECIO* atendiendo al de las variables que forman el modelo que nos ocupa, que facilitarán la

tarea de interpretar las variables que forman parte del modelo seleccionado, así como de los correspondientes parámetros estimados.

Recordando las propiedades del *Factor 1*, este parecía explicar las distintas variables relativas a valoraciones de los usuarios hacia los distintos hoteles, con lo cual, parece lógico que al aumentar estas, el precio aumente, hecho llevado a la ecuación de regresión en forma de un parámetro positivo y relevante multiplicando a esta variable.
Esta afirmación se sustenta con el aspecto presentado por la gráfica correspondiente, con una clara tendencia ascendente.

Puede llamar la atención los parámetros negativos estimados para las variables *CALIDADPRECIO* y *VALGLOBAL*. El papel que juegan estas variables en el modelo consiste en ajustar la información proporcionada por *Factor 1*. En general, el precio de un hotel tiende a subir a medida que las valoraciones recibidas por el mismo aumentan. No obstante, si nos fijamos en los gráficos correspondientes a las variables *CALIDADPRECIO* y *VALGLOBAL*, podemos ver ciertas bajadas en el precio a medida que aumentan estas valoraciones. La información proporcionada por la primera de estas variables no es más que una interpretación del cociente $calidad/precio$. Este cociente aumenta a medida que aumenta la calidad, pero también puede hacerlo si desciende el precio. Este es el caso de aquellos hoteles en los que, por un precio no muy alto, se ofrecen servicios muy buenos. En estas situaciones, el cliente queda plenamente satisfecho con el hotel, teniendo en cuenta la cantidad pagada, y por tanto, puntuará muy bien esta valoración concreta. Así, en los casos en que este cociente, o lo que es lo mismo, el valor de esta variable, fuera mayor que el del resto valoraciones, podemos asumir se debe a que el precio es razonablemente bajo, razón por la cual es necesaria la intervención en el modelo de esta variable acompañada de un parámetro negativo. Una explicación similar se podría aplicar al caso de la variable *VALGLOBAL*, ya que podemos asumir que un cliente queda satisfecho de forma general con un hotel si lo está con los servicios recibidos, y considera haber pagado un precio razonable por los mismos.
Además, respecto al ajuste relativo a esta variable, recordando la información contenida en la Tabla 1, es importante tener en cuenta que *VALGLOBAL* podría no estar lo suficientemente explicada por el *Factor 1*, o al menos, no tan bien como el resto de valoraciones.

Observando en conjunto las tres variables mencionadas hasta el momento, podemos observar que, pese a la naturaleza del *Factor 1* que pretendía explicar las distintas valoraciones, parece que aquellas relativas al precio del hotel necesitan un pequeño ajuste aparte, para cubrir algunas situaciones como las descritas anteriormente.

En cuanto al resto de variables que forman parte del modelo, todas ellas representan el precio promedio calculado con respecto a los diferentes niveles de las variables categóricas correspondientes. Así pues, es bastante natural que los parámetros estimados en estos casos sean positivos, y que a medida que aumenten estos promedios, aumente el precio estimado.

Fijándonos en las gráficas relativas a las variables categóricas correspondientes (Anexos Analíticos), podemos ver que el precio tiene en todos los casos la distribución esperada.

- Código Postal: para interpretar esta variable, hace falta un mínimo conocimiento de la ciudad de Madrid. Por ejemplo, uno de los códigos postales en los que el precio aumenta es el 28014, que abarca ciertas calles del centro de la ciudad, como lo son el Paseo del Prado, la calle Alcalá y la Plaza de las Cortes, entre otras, zona de la capital típicamente cara.
- Fecha: observar la gráfica de esta variable nos permite observar cómo aumenta el precio con la llegada de los fines de semana, como ocurre por ejemplo con los días 5 y 6, viernes y sábado respectivamente.
- Oferta: lógicamente, los precios se abaratan cuando están sujetos a ofertas (caso en que la variable *OFERTA* toma el valor 1).
- Estrellas: el precio es claramente más alto a medida que aumenta el número de estrellas.

4.3.3. Conclusión

Tras combinar y modificar la parrilla de parámetros y valores detallados en el punto anterior, se han obtenido 185 modelos, en una búsqueda que concluye con la elección de la estructura óptima.

La estimación del modelo se ha llevado a cabo llevó a partir de los datos de Entrenamiento. Por otra parte, la evaluación y comparación de estos modelos se ha desarrollado sobre el conjunto de datos de Validación. Como conclusión, se utiliza el conjunto de datos Test, con el objetivo de proporcionar una estimación insesgada del grado de acierto de nuestro modelo óptimo.

Sobre el último conjunto de datos mencionado, los datos *Test*, la precisión de nuestro modelo óptimo es muy acertada, con un error cuadrático medio de1113.86.En la Figura 7 se presenta un gráfico de cajas y bigotes, que representa el residuo calculado sobre los datos test (*precio-precio predicho*). Cada punto representa un *outlier*.

Como se puede observar, los valores tomados por los distintos residuos están centralizados alrededor del 0, excepto algunos casos extremos, lo cual nos permite afirmar que el precio predicho se ajusta bien al precio real, aún aplicando el modelo escogido como óptimo sobre un conjunto de datos diferente al empleado para su elaboración. Esto nos lleva a la conclusión de que el modelo seleccionado, además de garantizar una correcta predicción de la variable que nos ocupa, tiene una estructura que no depende estrechamente del conjunto de datos utilizado para su creación.

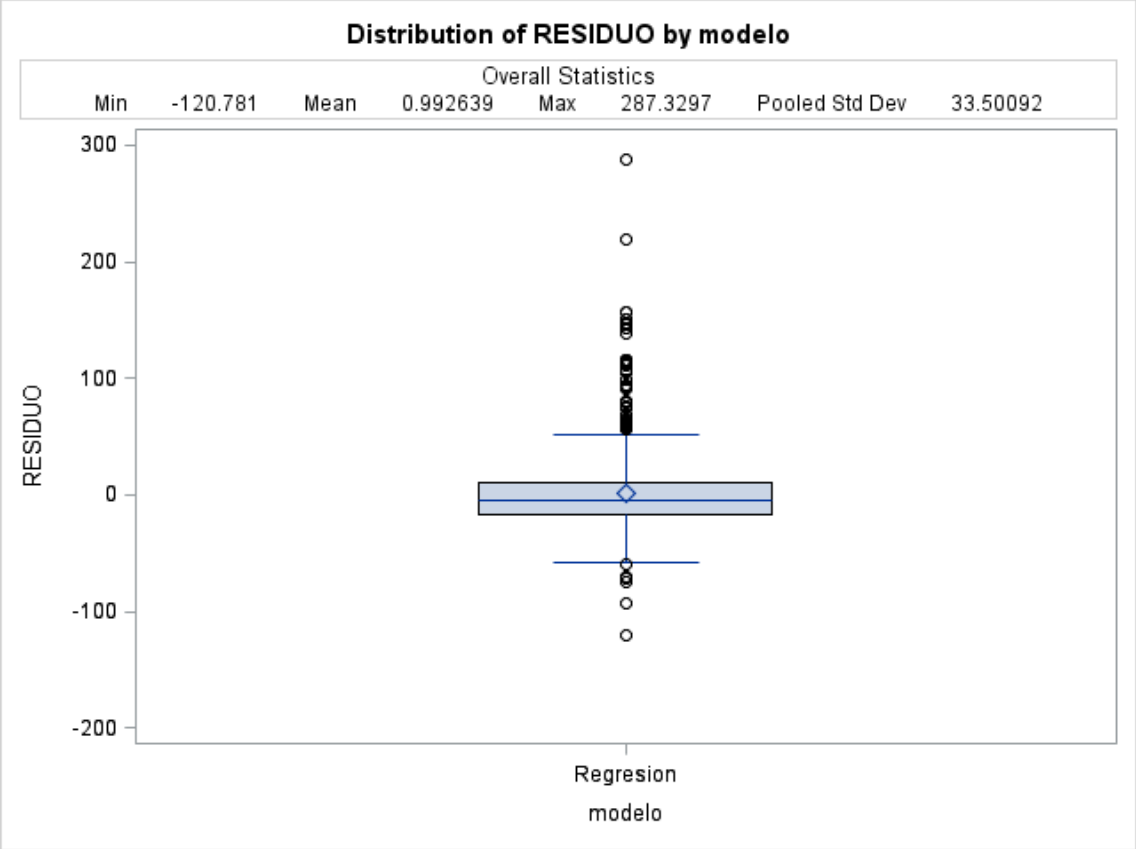

Figura 7. Distribución de los residuos calculados sobre los datos Test con el mejor modelo de regresión

4.4. Redes Neuronales

La aplicación de Redes Neuronales a la hora de encontrar un modelo para predecir la variable dependiente que nos ocupa, *Precio*, aportará una flexibilidad con la que no se contó en los procesos propios de la regresión. Por otra parte, nos ayudará en el tratamiento de las abundantes variables categóricas que componen la fuente de datos en uso, además de facilitar los procesos a llevar a cabo en un contexto en que la relación entre las variables no esté totalmente definida.

Es importante tener en cuenta que las Redes Neuronales tienden al sobreajuste (*Overfitting*), esto es, los modelos suelen estar demasiado ajustados a los datos utilizados para su construcción, funcionando relativamente mal para nuevos datos. Es entonces cuando cobra una importancia vital el uso de datos de validación y el posterior test.

Además, resultará primordial evitar que el azar no intervenga demasiado en las decisiones. Para ello, los procesos de partición, modelado y comparación se llevarán a cabo utilizando diferentes semillas para la partición aleatoria, lo cual nos permitirá encontrar un punto de coherencia entre los modelos.

4.4.1. Obtención de modelos. Búsqueda del modelo óptimo.

La búsqueda del mejor modelo de Redes Neuronales se lleva a cabo de forma empírica en un proceso prueba-error, consistente en la obtención y comparación de distintos modelos mediante la aplicación de distintos métodos de selección y ajuste de parámetros. Con ello, se han obtenido un total de 503 modelos distintos, todos ellos susceptibles de variaciones. Es importante recordar que esta búsqueda se llevará a cabo sobre el conjunto de datos de entrenamiento o aprendizaje.

En general, las distintas variaciones que se han obtenido se han llevado a cabo modificando los siguientes parámetros:

- Conjunto de **variables** a utilizar en la red. Se construyen modelos utilizando, por una parte, todo el conjunto de variables en uso de la base de datos, y por otro, el conjunto de variables que forman parte del modelo óptimo de regresión, detallado en el epígrafe anterior, a saber, *FACTOR1*, *CALIDADPRECIO*, *UBICACIÓN*, *VALGLOBAL*, *CPPROM*, *FECHAPROM*, *OFERTAPROM* y *ESTRELLASPROM*.
- **Algoritmo de Optimización (***BadPropagation* y *Levmar*).
- **Función de Combinación** (Lineal y Aditiva)
- **Función de Activación** (Tangente Hiperbólica, Elliot, Exponencial, Arco Tangente, Lineal, Cuadrado, Tangente y Logarítmica).
- Número de **nodos** en la capa oculta (1,2,…,30). La complejidad computacional de añadir más de una capa oculta, sumada al hecho de que nuestra fuente de datos no es de alta complejidad, nos han llevado a llevar a cabo pruebas con una sola capa oculta, y con un máximo de 30 nodos.
- Modificación de las **semillas de aleatorización**, con el objetivo de controlar el efecto del azar sobre los resultados obtenidos.
- Máximo de **iteraciones** permitidas.
- Aplicación de ***Early Stopping*** (utilizaremos esta técnica para intentar regularizar el sobreajuste propio de los métodos en uso).

A la hora de comparar los modelos obtenidos, nos fijaremos, esencialmente, en el error cuadrático medio.

Tras obtener los diferentes modelos combinando la parrilla de valores descrita, obtenemos un modelo susceptible de ser elegido como óptimo, cuyo error cuadrado medio calculado sobre datos de validación, y parámetros son:

- **Modelo (**Error Cuadrado Medio: 482.77)
    - Conjunto de variables: variables del mejor modelo de regresión.
    - Algoritmo de Optimización: *Levmar*.
    - Función de Combinación: Lineal.
    - Función de Activación:Tangente Hiperbólica.
    - Número de nodos ocultos: 17.
    - Número de capas ocultas: 1.
    - Máximo de iteraciones permitidas:80
    - Uso de *Early Stopping*: sí

El número de parámetros a estimar en una Red Neuronal con una sola capa y una variable output se calcula con la fórmula $h(k+1)+h+1$, siendo h el número de nodos ocultos, y k el número de variables de entrada. Así, para el modelo mencionado, habría que estimar 171 parámetros.

A continuación, en las figuras 8 y 9, podemos observar cómo varía el comportamiento del modelo a la hora de predecir, con o sin la aplicación de la técnica de *Early Stopping*.

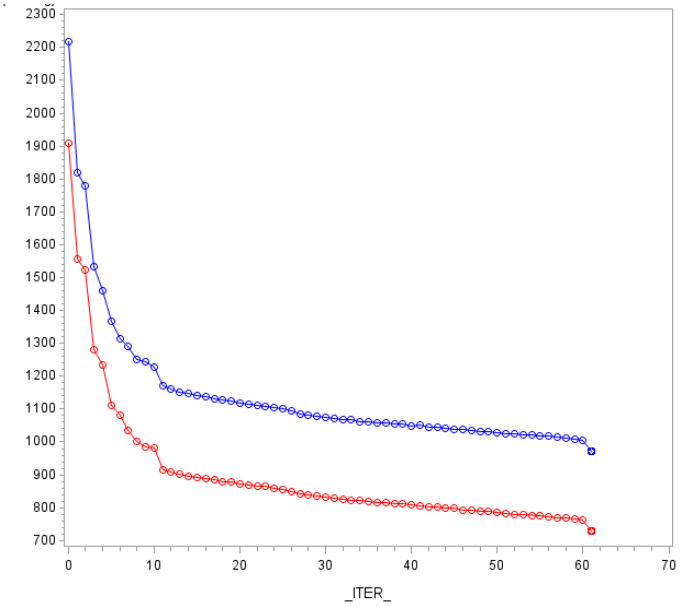

Figura 8. Sin Early Stopping

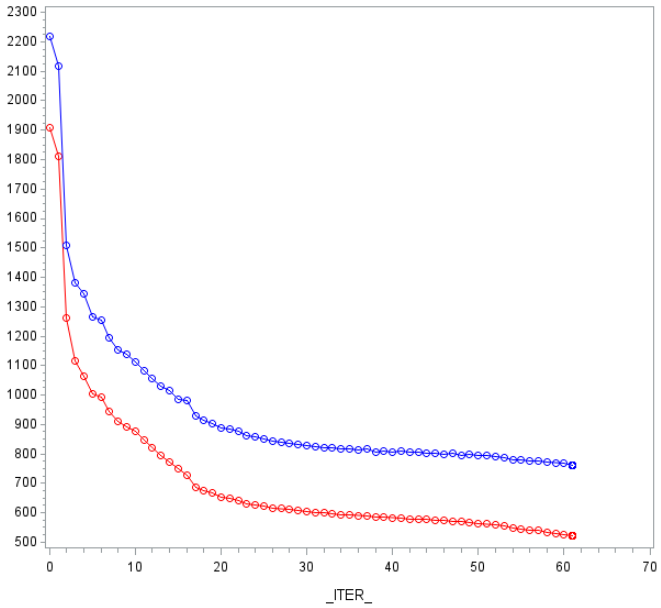

Figura 9. Con Early Stopping

Como se puede apreciar en los gráficos, la aplicación de *Early Stopping* mejora en cierta medida la red obtenida. Además, este proceso optimiza el error cuadrado medio del modelo. Si calculamos esta medida de ajuste sobre los datos de validación, aplicando la estructura seleccionada como candidata a óptima pero sin utilizar *Early Stopping*, obtenemos 488.24, que es peor que el error cuadrado medio obtenido con la aplicación de la misma red neuronal, combinada con el uso de *Early Stopping* (482.77).

Con el objetivo de evaluar la "importancia" de cada una de las variables independientes dentro de la Red Neuronal, y puesto que este resultado no se puede conseguir con la aplicación del método *Proc Neural*, se procede al cálculo de un total de 8 redes más, repitiendo la estructura de la que en este punto es la mejor red obtenida, eliminando en cada uno de los nuevos modelos una única variable explicativa. Después, se calculan las medidas de ajuste de las nuevas redes, y se comparan con las resultantes de la red "inicial".

VARIABLE QUITADA	ERROR CUADRADO MEDIO	Nº DE PARÁMETROS A ESTIMAR
FACTOR 1	546.42	154
CALIDADPRECIO	562.20	154
UBICACIÓN	537.01	154
VALGLOBAL	574.76	154
CPPROM	566.57	154
FECHAPROM	585.19	154
OFERTAPROM	515.14	154
ESTRELLASPROM	504.74	154
Red Original	**482.77**	**171**

Tabla 2. Redes Auxiliares

Como se puede observar en la Tabla 2, ninguna de las nuevas redes obtenidas con la eliminación de las variables independientes una a una mejora a la red de partida (el error cuadrático medio está calculado sobre los datos de validación).

Así, estamos en condiciones de afirmar que el modelo de Red Neuronal óptimo se construye con los siguientes parámetros:

- **Modelo (**Error Cuadrado Medio: 482.77)
    - Conjunto de variables: variables del mejor modelo de regresión.
    - Algoritmo de Optimización: *Levmar*.
    - Función de Activación: Tangente Hiperbólica.
    - Número de nodos ocultos: 17.
    - Número de capas ocultas: 1.
    - Máximo de iteraciones permitidas: 80
    - Uso de *Early Stopping*: sí

4.4.2. Interpretación del modelo.

Antes de continuar con el desarrollo de este punto, cabe hacer un paréntesis para definir una propiedad de las variables, llamada importancia. Así, diremos que la **importancia de una variable independiente** es una medida que indica cuánto cambia el valor pronosticado por el modelo de Redes Neuronales para diferentes valores de la variable independiente. La importancia normalizada es el resultado de los valores de importancia divididos por el valor de importancia mayor, expresados como porcentajes.
En nuestro caso, la importancia se establecerá en función de la variabilidad del error cuadrado medio, y queda recogida en el siguiente gráfico de barras.

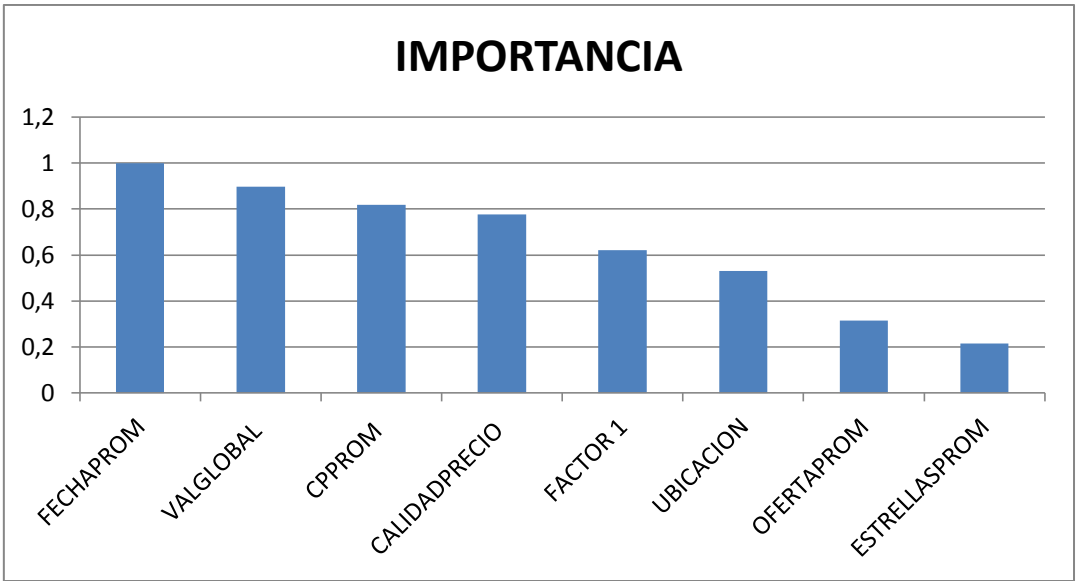

Figura 10. Importancia de las variables en la Red Neuronal óptima

Como se puede observar, a la hora de construir el modelo, la variable más importante es *FECHAPROM*, que es un indicativo del momento temporal en que se va a llevar a cabo la reserva. No obstante, como muestra el gráfico, todas las variables que forman parte de la red neuronal tienen una importancia latente en el modelo, hecho reflejado también en la Tabla 2, que nos muestra cómo empeora el modelo óptimo si se le elimina alguna de las variables.

4.4.3. Conclusión.

Tras llevar a cabo una búsqueda exhaustiva de la mejor estructura de Redes Neuronales, durante la cual se han creado y comparado más de 500 modelos, resultantes de la combinación de distintos valores para una serie de parámetros, y una vez fijado el mejor modelo, se ha intentado reducir la dimensión de las variables explicativas, aún sin éxito. Esto nos lleva a afirmar fehacientemente que nuestro modelo de Redes Neuronales está formado por un conjunto mínimo de variables. La estimación del modelo se ha desarrollado sobre los datos de Entrenamiento, mientras que la evaluación y comparación de los distintos modelos se ha llevado a cabo sobre los datos de Validación. Para concluir con este aparto, se proporcionará una estimación insesgada del grado de acierto del modelo establecido como óptimo, para lo cual se aplicará la estructura de Redes Neuronales óptima sobre el conjunto de datos Test, obteniendo un error cuadrado medio de 828.39.

En la Figura11 se puede observar un gráfico de cajas y bigotes, en el cual podemos observar el comportamiento de los residuos correspondientes a la predicción de la variable dependiente que nos ocupa sobre el conjunto de datos Test, utilizando la estructura de Redes Neuronales establecida como óptima.

Exceptuando los *outliers*, representados por los puntos blancos, podemos ver que el rango de valores abarcado por los residuos es más pequeño que en el caso de regresión.

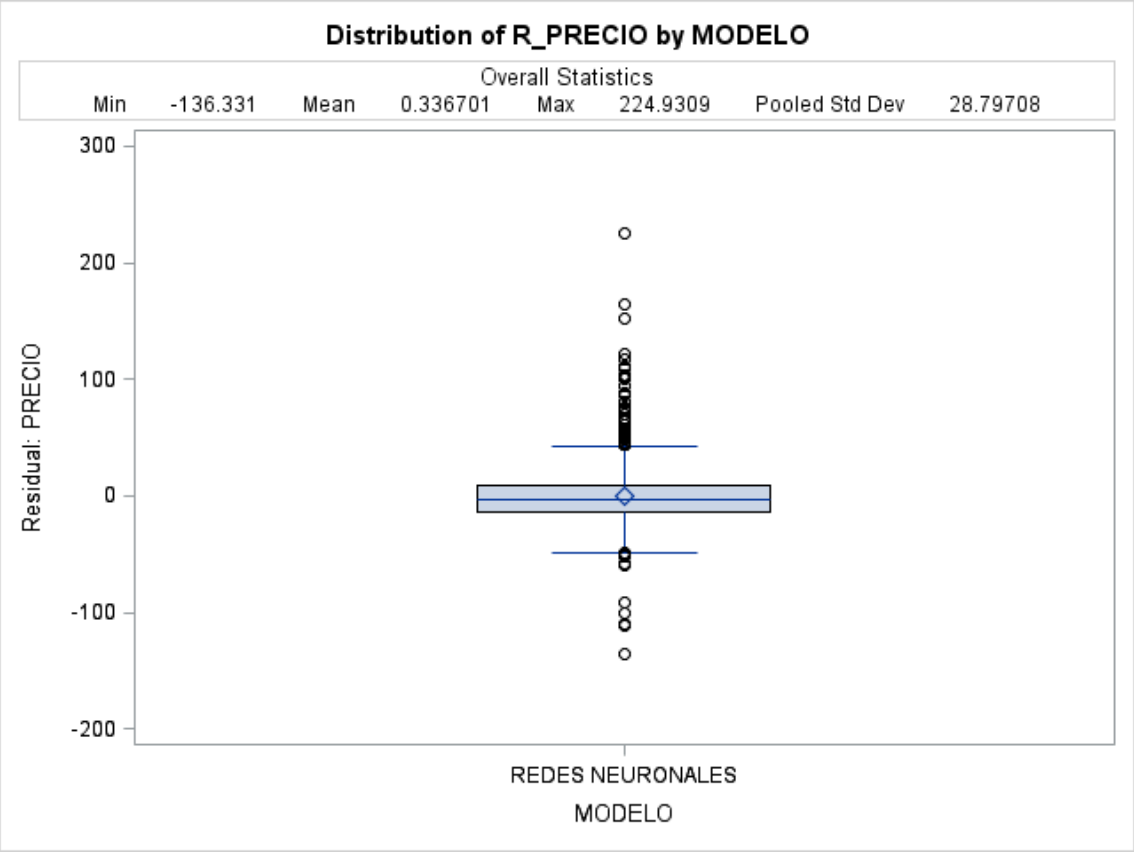

Figura 11. Distribución de los residuos calculados sobre datos Test con el mejor modelo de Redes Neuronales

Si interpretamos la información recogida en este gráfico, en el cual podemos ver que todos los valores de los residuos se agrupan en torno al 0, podemos afirmar que nuestro modelo ha llevado a cabo una buena predicción, en la que el dato predicho se aproxima muy acertadamente al dato real. El hecho de haber llevado a cabo esta prueba sobre el conjunto de datos Test obteniendo un resultado satisfactorio, nos ofrece garantías de que esta estructura funcionará, en general, sobre cualquier conjunto de datos, ya que el conjunto de datos sobre el cual se ha construido el modelo no es altamente determinante.

4.5. Random Forest

Random Forest es una técnica de agregación que mejora la predicción con respecto a otros métodos mediante la inclusión de aleatorización en la construcción de las distintas estructuras.

En las siguientes páginas se desarrollará el proceso de búsqueda, elección y comprensión de la mejor estructura arbórea construida con la aplicación de este método.

4.5.1. Obtención de modelos. Búsqueda del modelo óptimo

La construcción del mejor modelo *Random Forest* se lleva a cabo de forma empírica en un proceso de prueba-error en el cual se crea una parrilla de valores para los diferentes parámetros propios de esta técnica, calculándose el error cuadrático medio para cada modelo, así como el número de reglas utilizadas para su construcción.

Los distintos modelos se han obtenido mediante la variación de los parámetros que se presentan a continuación:

- **Variables a utilizar**. Se harán pruebas utilizando distintas combinaciones de las variables disponibles.
- ***p*-valor** necesario para generar una regla de división. Se han hecho pruebas para los valores 0.1, 0.07, 0.05, 0.03 y 0.01.
- **Tamaño de hoja mínimo**, considerando los valores 5, 10, 50, 755, 100, 150 y 200.
- **Máximo número de árboles** a crear, probando con 5, 10, 15, 20, 25, 40 y 50 como número máximo.
- **Número** de **variables** a tener en cuenta uno de los nodos de los diferentes arboles, buscando modelos con 3, 6, 9, 12, 15 y 18 variables .
- **Porcentaje** de la población que se **muestrea** en la construcción de cada árbol (valor que dejamos fijo en un 68%.)

Combinando los valores mencionados para los distintos parámetros, se obtiene un total de 1980 modelos. Es importante mencionar que, en general, aparentemente los modelos creados se muestran bastante estables ante las variaciones propuestas para los parámetros estructurales "p-valor" y "Número máximo de árboles a crear".
Así, con un error cuadrático medio de 405.24 calculado sobre datos de validación, y 2908 reglas necesarias para su creación, el mejor modelo de *Random Forest* obtenido se consigue dando los siguientes valores a los parámetros:

- **Variables a utilizar:** *VALGLOBAL UBICACION CALIDADPRECIO FACTOR1 OFERTA FECHA ESTRELLAS ZONA CODPOSTAL DIA FINDE COMENTARIOS*
- ***p*-valor**: 0.07.

- **Tamaño de hoja mínimo:** 10.
- **Máximo número de árboles a crear**: 20.
- **Número de variables**: 6.

A la hora de escoger un modelo como óptimo, es importante tener en cuenta que los valores asignados a los parámetros que lo conforman no deben ser extremo inferior ni superior del rango de posibles valores asignados a cada parámetro. En caso de llegar a esta situación, habría que seguir buscando, ampliando el rango de valores a asignar a cada parámetro. Por ejemplo, si la primera parrilla de valores a probar asignada al parámetro *tamaño mínimo de hojas* fuera (10, 15, 50), y el resultado obtenido nos dijera que el modelo óptimo se construye dando el valor 50 a este parámetro, habría que realizar más pruebas, asignando más valores (superiores a 50) a dicho parámetro, ya que no podemos garantizar que el modelo óptimo necesite esta cantidad, siendo probable que el número ideal sea superior a este.

En nuestro caso, ninguno de los valores asignados a los parámetros del modelo que minimiza el error cuadrático medio y el número de reglas necesarias para su creación, sufre esta situación, por lo que, podemos afirmar que, aún ampliando inferior y superiormente el rango de valores, el modelo seleccionado es óptimo.

4.5.2. Interpretación del modelo.

Una vez establecido el modelo óptimo de *Random Forest*, se procede a interpretar los valores establecidos para los parámetros que lo conforman.
En este contexto, y puesto que los árboles de decisión no se ven afectados por relaciones lineales entre variables, y facilitan el tratamiento de las variables, especialmente de las categóricas, las primeras pruebas se llevaron a cabo utilizando todas las variables de la base, para, finalmente, quedarnos con el conjunto presentado en el epígrafe anterior.

Como se puede observar, las variables de naturaleza cualitativa intervienen como tal, sin utilizar las correspondientes cuantitativas (como ocurría en los procesos de regresión y redes neuronales). Además, se puede observar que prácticamente todas las variable de la base de datos tiene un papel a la hora de predecir aplicando el método que nos ocupa.
En este caso, establecemos que el máximo número de árboles a crear sea 20, cantidad que parece bastante razonable, en el contexto que nos ocupa, donde nuestra base de datos no supera las 10.000 observaciones. El tamaño de las hojas mínimo nos garantiza que no se crearán hojas con menos de 10 observaciones, lo cual nos permite afirmar que, aunque nuestro problema inicial se irá simplificando en un conjunto problemas más simples (esta es la filosofía de los árboles de decisión, y, en concreto, del *Random Forest*), nunca llegaremos a casos triviales.

4.5.3. Conclusión

Se ha llevado a cabo una búsqueda exhaustiva de la mejor estructura de *RandomForest*. En este proceso, se han creado un total de 1980 modelos diferentes, resultado de combinar un conjunto de parámetros con un amplio rango de posibles valores.

La estimación del modelo, llevada a cabo a partir de los datos de entrenamiento, es posteriormente evaluada sobre el conjunto de datos de validación, para, finalmente, proporcionar una estimación insesgada del grado de cierto del modelo seleccionado como óptimo, para lo cual se utilizará el conjunto de datos Test. Tras aplicar el modelo seleccionado sobre el conjunto de datos mencionados, obtenemos un error cuadrado medio de 582.62, lo cual nos indica que, efectivamente, nuestro modelo se ajusta a unos datos no usados para su construcción de una forma bastante acertada.
A continuación se presenta el gráfico de cajas y bigotes que representa los residuos correspondientes a la predicción de nuestra variable dependiente usando el modelo óptimo de *Random Forest*¸ sobre el conjunto de datos Test.

En el mismo se puede observar que los valores de los residuos están, en general, centralizados en torno al 0, lo cual quiere decir que el error predicho con la aplicación de esta técnica se ajusta muy acertadamente al precio real.

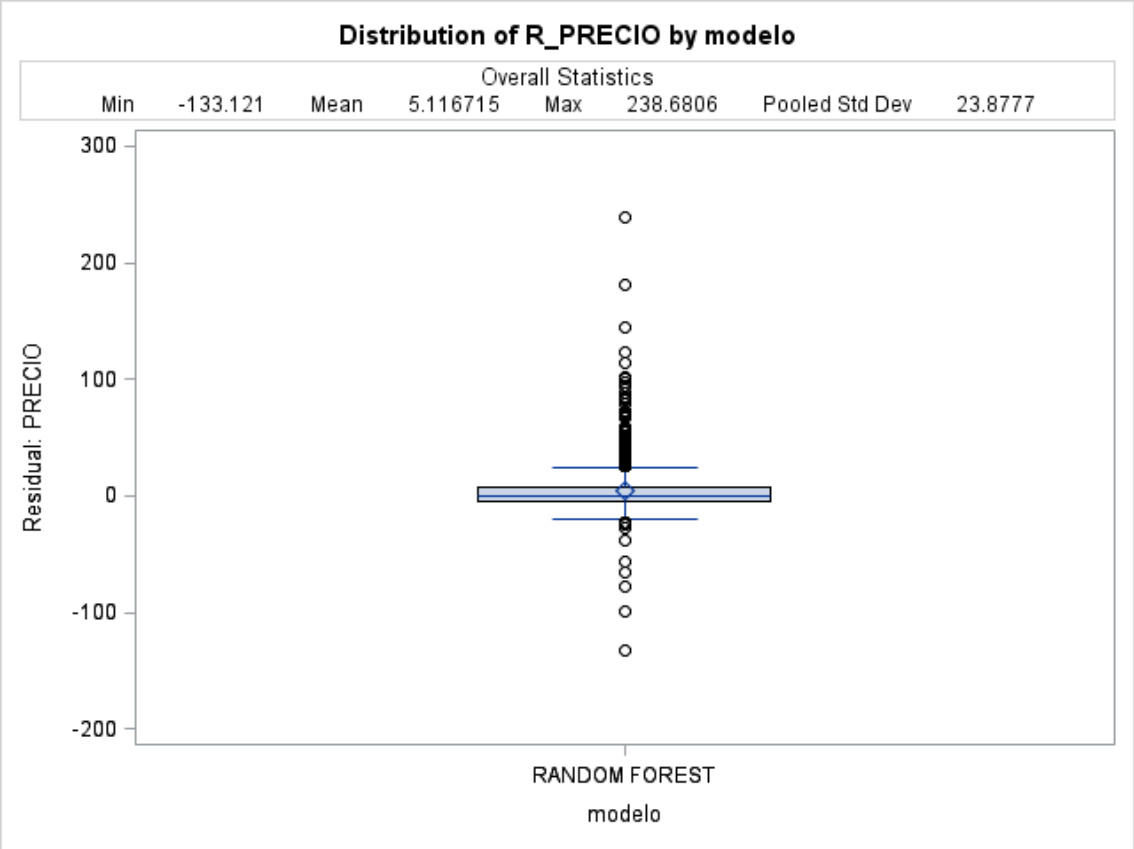

Figura 12. Distribución de los residuos calculados sobre los datos test con el mejor modelo de Random Forest.

4.6. Gradient Boosting

El método que ahora nos ocupa, *Gradiend Boosting*¸ es una técnica que pretende mejorar las estructuras de árboles de decisión. En las siguientes páginas se detalla el proceso de búsqueda, elección y comprensión de la estructura óptima obtenida con la aplicación de este método.

4.6.1. Obtención de modelos. Búsqueda del modelo óptimo

La búsqueda del mejor modelo de *Gradient Boosting* se lleva a cabo de forma empírica en un proceso prueba-error, consistente en la obtención y comparación de distintos modelos mediante la aplicación de distintos métodos de selección y ajuste de parámetros. Con ello, se han obtenido un total de 3906 modelos distintos, obtenidos sobre los datos de entrenamiento.
Para ello, a cada uno de los parámetros que intervienen en el proceso, se le ha asignado un conjunto de valores con los que hacer distintas pruebas y combinaciones. Así, la parrilla de valores para los siguientes parámetros es la siguiente:

- **Variables a utilizar**. Se harán pruebas utilizando distintas combinaciones de las variables disponibles.
- **Tamaño de hoja mínimo:** se harán pruebas asignando a este parámetro los valores 10, 25, 50, 75, 100, 125, 200, 300, 400, 500 y 600.
- **Máximo número de árboles** a crear, asignando a este parámetro la parrilla de valores 125, 100, 75, 50,25 y 10.
- **Parámetro de Regularización**: a este parámetro se le asignarán los valores 0.01, 0.03, 0.05, 0.07, 0.09, 1.1, 1.3, 1.5 y 1.7.
- **Iteraciones**se llevarán a cabo distintas pruebas, estableciendo como posibles valores para el número máximo de iteraciones 10, 25, 50, 75, 100, 125, 125 y 175.
- **Profundidad máxima del árbol:** se harán pruebas asignando a este parámetro los valores 25 y 50, que es el máximo permitido.

Al igual que sucede al aplicar el algoritmo *Random* Forest, a la hora de escoger un modelo como óptimo, es importante tener en cuenta que los valores asignados a los parámetros que lo conforman no deben ser extremo inferior ni superior del rango de posibles valores asignados a cada parámetro. En caso de llegar a esta situación, habría que seguir buscando, ampliando el rango de valores a asignar a cada parámetro.
En nuestro caso, tras probar con toda la parrilla de valores comentada obteniendo los distintos modelos, con un error cuadrático medio de 206.10 calculado sobre los datos validación, y 1272 reglas necesarias para su creación, seleccionamos el siguiente modelo como candidato a óptimo.

- **Variables a utilizar:** *OFERTA FECHA ESTRELLAS ZONA CODPOSTAL COMENTARIOS DIA FINDE VALGLOBAL UBICACION CALIDADPRECIO WIFI FACTOR*1
- **Tamaño de hoja mínimo**: 100

- **Máximo número de árboles a crear**: 25
- **Parámetro de regularización**: 0.09
- **Iteraciones**: 100
- **Profundidad máxima del árbol**: 25

Si bien este no es el modelo que minimiza el error cuadrático medio, está sólo a un 4% de distancia (en cuanto a eficacia se refiere) del modelo que minimiza este valor. No obstante, para crear esta estructura hacen falta más de 5000 reglas, lo cual constituye una razón de peso para considerar mejor el modelo candidato a óptimo, que podría crearse con sólo 1272 reglas. Además, si observamos el resto de parámetros, no toman valores muy grandes, lo cual garantiza a nuestro modelo cierta robustez y estabilidad.

Por tanto, podemos afirmar que, con un error cuadrático medio de 206.10 y 1272 reglas necesarias para su construcción, el modelo óptimo obtenido con la aplicación del algoritmo *Gradient Boosintg* se obtiene dando los siguientes valores a los parámetros:

- Variables a utilizar**:** *OFERTA FECHA ESTRELLAS ZONA CODPOSTAL COMENTARIOS DIA FINDE VALGLOBAL UBICACION CALIDADPRECIO WIFI FACTOR*1
- Tamaño de hoja mínimo: 100
- Máximo número de árboles: 25
- Parámetro de Regularización: 0.09
- Iteraciones: 100
- Profundidad máxima del árbol :25

4.6.2. Interpretación del modelo

Una vez calculado el modelo óptimo obtenido aplicando la técnica *Gradient Boosting*, se procede a hacer una interpretación de los valores establecidos para sus parámetros.

Al igual que ocurre en el caso de *Random Forest*, la naturaleza del algoritmo *Gradient Boosting* hace que no se vea afectado por relaciones lineales entre las variables, y facilita el tratamiento de variables. Así, en el modelo seleccionado como óptimo intervienen la mayoría de las variables disponibles. Además, aquellas que inicialmente son de naturaleza cualitativa, son introducidas en el modelo como tal, y no utilizando la correspondiente cuantitativa, tal y como se detalla en los procesos de regresión y redes neuronales.

Continuamos analizando el número de reglas necesarias para su creación. Si bien es cierto es que la cantidad establecida no es la más pequeña de todas las probadas, sí podemos afirmar que, de entre los valores asignados a este parámetro que garantizan un error cuadrático medio aceptable, es la más pequeña. Así, todas las estructuras obtenidas con menos de 1000 reglas vienen acompañadas de un error cuadrático medio superior a 800. Por tanto, podemos afirmar que en este caso compensa aumentar el número de reglas, debido a la brusca reducción de error.

Por otra parte, si nos fijamos en el tamaño de hojas mínimo, 100, nos garantizará que no se formarán hojas con menos de esta cantidad de observaciones. Partiendo de la cantidad de hoteles en uso, esta cantidad podemos afirmar que este parámetro garantizará la obtención de un modelo robusto.
En cuanto a la profundidad máxima del árbol, está fijada en 25. Esta cantidad, no demasiado grande, tiene sentido en el contexto que nos ocupa, ya que la base de datos en uso tiene menos de 10.000 datos, y quizás, establecer una profundidad mayor que esta nos llevaría a la creación de árboles demasiado sencillos, con muy pocas observaciones.

4.6.3. Conclusión

Se ha llevado a cabo una búsqueda exhaustiva de la mejor estructura de *Gradient Boosting*. En este proceso, se han creado un total de 3906 modelos diferentes, resultado de combinar un conjunto de parámetros con un amplio rango de posibles valores.
La estimación del modelo, llevada a cabo a partir de los datos de entrenamiento, es posteriormente evaluada sobre el conjunto de datos de validación, para, finalmente, proporcionar una estimación insesgada del grado de cierto del modelo seleccionado como óptimo, para lo cual se utilizará el conjunto de datos Test. Utilizando dicho conjunto de datos, obtenemos, con la aplicación del modelo óptimo, un error cuadrado medio de 319.76, lo cual nos permite afirmar que nuestro modelo se ajusta a los datos en uso a la hora de hacer una predicción.

Si observamos el gráfico de cajas y bigotes de la Figura 13, en el cual están representados los residuos obtenidos al predecir el precio sobre el conjunto de datos Test aplicando el modelo de *Gradient Boosting* seleccionado como óptimo, se puede observar que, en general, todos los valores están centralizados en torno a 0, lo cual quiere decir que $\text{precio} - \text{precio\_predicho} \cong 0$. Esto nos sirve como indicativo de que la estructura elegida ha tenido un funcionamiento correcto.

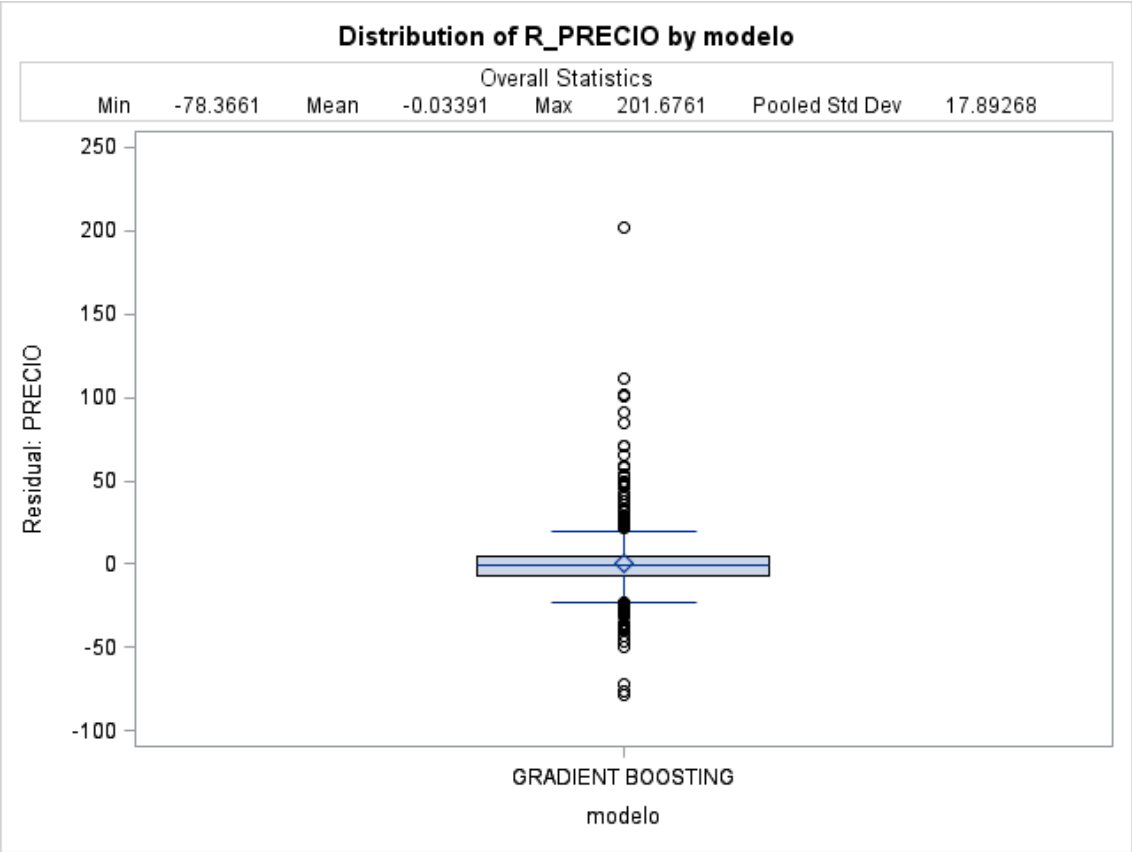

Figura 13. Distribución de los residuos calculados sobre datos test con el mejor modelo de Gradient Boosting.

4.7. Elección del modelo óptimo

Llegados a este punto en que ha sido seleccionada una estructura óptima para cada técnica utilizada, es el momento de elegir el mejor modelo de entre todos ellos.
Para ello, nos fijaremos, en el error cuadrático medio calculado sobre los datos de validación para cada uno de estos modelos. En la Tabla3podemos ver un resumen de los mismos.

ESTRUCTURA	ERROR CUADRÁTICO MEDIO
REGRESIÓN	824.15
REDES NEURONALES	482.77
RANDOM FOREST	405.24
GRADIENT BOOSTING	206.10

Tabla 3. Modelos óptimos

Como podemos ver, el modelo que minimiza el error cuadrático medio es el de *Gradient Boosting*, seguido por el de *Random Forest*.
No obstante, si recordamos las estructuras de ambos modelos, podemos ver que la de *Random Forest* es algo más sencilla que la de *Gradient Boosting*, lo cual podría garantizar algo más de estabilidad al modelo.

Random Forest (Error Cuadrático Medio: validación, 405.24; test, 582.62))

- Variables a utilizar**:** $VALGLOBAL\ \ UBICACION\ \ CALIDADPRECIO\ \ FACTOR1\ \ OFERTA\ \ FECHA\ \ ESTRELLAS\ \ ZONA\ \ CODPOSTAL\ \ DIA\ \ FINDE\ COMENTARIOS$
- p-valor : 0.07.
- Tamaño de hoja mínimo: 10.
- Máximo número de árboles a crear: 20.
- Número de variables: 6.

Gradient Boosting (Error Cuadrático Medio: validación 206.10; test , 319.76)

- Variables a utilizar**:** $OFERTA\ \ FECHA\ \ ESTRELLAS\ \ ZONA\ \ CODPOSTAL\ \ COMENTARIOS\ \ DIA\ \ FINDE\ \ VALGLOBAL\ \ UBICACION\ \ CALIDADPRECIO\ \ WIFI\ \ FACTOR1$
- Número de reglas: 1272
- Tamaño de hoja mínimo: 100
- Máximo número de árboles: 25
- Parámetro de Regularización: 0.09
- Iteraciones: 100
- Profundidad máxima del árbol: 25

Por otra parte, en la Figura 14 podemos ver la distribución de los residuos de ambos modelos, calculados sobre el conjunto de datos test. Podemos ver que en ambos casos se distribuyen muy uniformemente y en un rango de valores bastante pequeño (excepto *outliers*, que parecen más abundantes en el caso de *Random Forest*).

Finalmente, tras calibrar el coste de crear cada estructura, y los beneficios que esta aporta a la eficiencia a la hora de predecir, se establece que el modelo óptimo de predicción obtenido es el de *Gradient Boosting*, con los siguientes parámetros:

Gradient Boosting (Error Cuadrático Medio: validación 206.10; test , 319.76)

- Variables a utilizar**:** $OFERTA\ \ FECHA\ \ ESTRELLAS\ \ ZONA\ \ CODPOSTAL\ \ COMENTARIOS\ \ DIA\ \ FINDE\ \ VALGLOBAL\ \ UBICACION\ \ CALIDADPRECIO\ \ WIFI\ \ FACTOR1$
- Número de reglas: 1272
- Tamaño de hoja mínimo: 100
- Máximo número de árboles: 25
- Parámetro de Regularización: 0.09
- Iteraciones: 100
- Profundidad máxima del árbol: 25

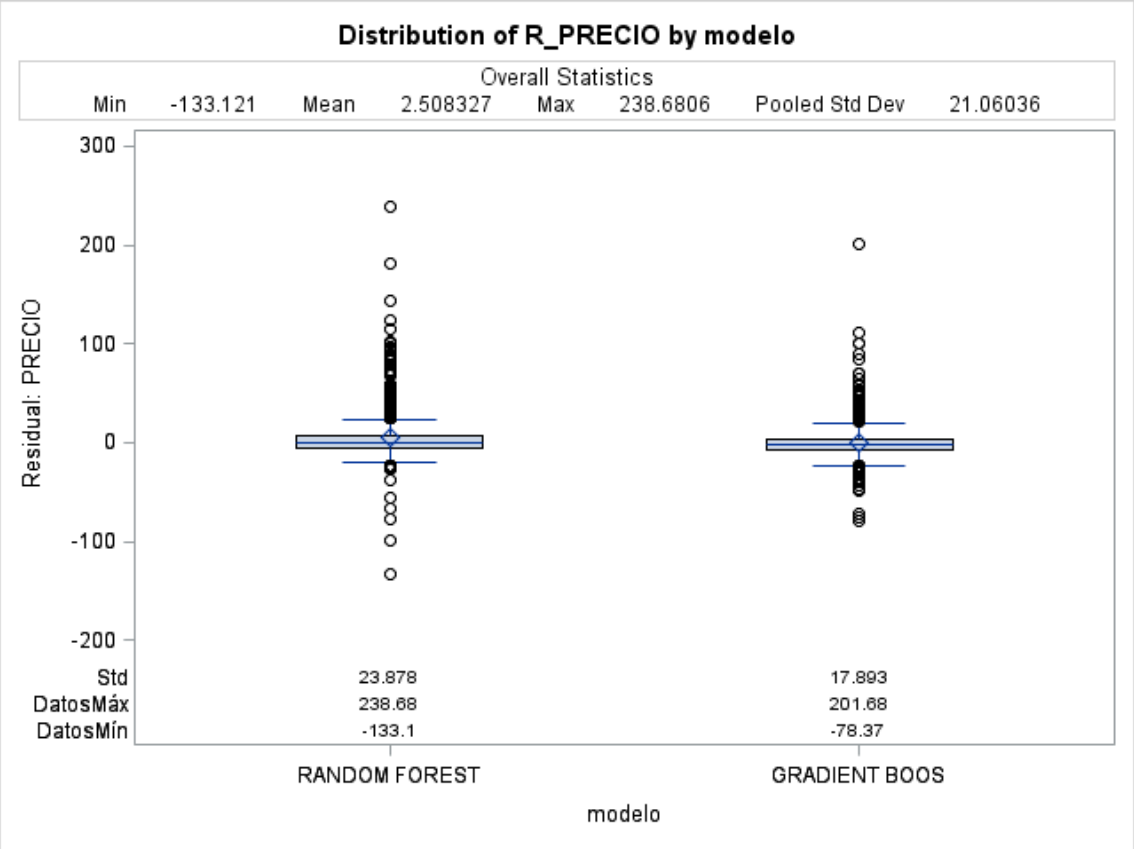

Figura 14. Residuos obtenidos con los modelos de Random Forest y Boosting sobre datos test

5. Análisis de Competencia Empresarial

En este punto se llevará a cabo un análisis de competencia empresarial en un proceso consistente en relacionar una empresa con su entorno o competencia para, posteriormente, determinar si el hotel escogido está dentro o fuera de su mercado, esto es, si los precios que establece se ajustan a los marcados por los centros que le hacen competencia.

Para ello, se combinarán dos técnicas multivariantes, como lo son el Análisis Factorial (detallado en el epígrafe 4, correspondiente a modelos de predicción) y el Análisis *Clúster*, con procesos y técnicas propias del análisis de competencia empresarial.

El hotel escogido para ejemplificar el proceso es el *Hotel Europa,* de 3 estrellas, y situado en la zona centro de la ciudad. Se ha escogido esta empresa por ser la primera en aparecer en la lista que nos ocupa, pero del mismo modo se podría haber tomado cualquier otra de todas las que integran la lista que nos ocupa.

5.1. Análisis Cluster

Tomando los factores obtenidos en el proceso llevado a cabo correspondiente a análisis factorial, procederemos a realizar un análisis *cluster* mediante el cual se clasificarán los hoteles en distintos grupos, en base a semejanzas y diferencias entre sus perfiles. Para crear estos grupos, nos basaremos en la información obtenida de los factores, y de la variable estrella (como se menciona en páginas anteriores, este detalle que a priori resulta importante a la hora de establecer las características de un hotel y su competencia, no queda explicado por los factores en uso). Así, para cada hotel, su competencia quedará establecida por los establecimientos que formen parte de su mismo grupo. El objetivo perseguido en este punto es conseguir clasificar las distintas observaciones en grupos que tengan las siguientes propiedades:

- Cada grupo debe ser homogéneo con respecto a las variables utilizadas para su formación.
- Los grupos deben ser lo más distintos posible unos de otros.

Es importante tener en cuenta que, a priori, la composición de los grupos es desconocida.

Por la interpretación dada en el punto anterior a los factores obtenidos, en general estos grupos se formarán atendiendo a las valoraciones del hotel, a la zona en que se encuentren, el momento temporal a realizar la reserva, y añadiremos las estrellas de cada centro.
Se harán varias pruebas en las que la competencia de cada hotel, (o grupo creado), variará atendiendo a los factores utilizados para la creación de los *cluster*.

En general, el proceso a seguir en este tipo de análisis se basa en:

- ***Clúster* Jerárquico** de carácter divisivo. Es el primer paso del proceso, y se lleva a cabo en un punto en que todas las observaciones comienzan en el mismo grupo, y se van realizando divisiones, aún sin saber cuál es la cantidad exacta

óptima de grupos a crear en cada caso. Esta técnica nos proveerá de gráficos en forma de dendrograma, en los cuales los datos quedarán organizados en subcategorías que se van subdividiendo hasta llegar al nivel de detalle deseado. Este tipo de representación nos permitirá apreciar claramente las relaciones de agrupación entre nuestros datos, lo cual será de utilidad a la hora de conocer para cada hotel su competencia.

Este proceso será llevado a cabo utilizando como base los resultados obtenidos al realizar distintas pruebas con el proceso de SAS *proc cluster*. Se harán intentos con los métodos *Average*(este método utiliza como medida de vinculación entre observaciones la técnica de agrupamiento de pares no ponderados, con el uso de promedios aritméticos, donde se establece que la distancia entre dos grupos es la distancia media entre todos los pares de objetos de ambos grupos), y *Ward* (este método aplica la técnica de la mínima varianza de Ward, donde el criterio para la elección del par de *clusters* a mezclar en cada paso se basa en el valor óptimo de una función objetivo, que en nuestro caso será el error de la suma de los cuadrados o varianza).

- ***Clúster* no Jerárquico.** Una vez conocido el número concreto de grupos a crear, procedemos a llevar a cabo un proceso de *clúster*, en este caso, no jerárquico. Apoyándonos en el proceso de SAS *proc fastclous*¸ esta técnica nos permitirá crear grupos homogéneos e independientes entre sí, tantos como se hayan establecido en el paso jerárquico previo.

5.2. Resultados obtenidos.

A continuación se mostrarán los dendrogramas correspondientes a los distintos resultados obtenidos. Como se menciona en el punto anterior, se busca poder agrupar los hoteles, atendiendo a distintos aspectos, para lo cual, se llevarán a cabo varios procesos de análisis *clúster*, cada uno de los cuales incluirá una combinación distinta de los factores obtenidos en los pasos previos, haciendo variaciones también con la intervención de la variable *ESTRELLAS*.

5.2.1. Factor 1, Factor 2, Factor 3 y Estrellas

Los primeros pasos en cuanto a análisis clúster se refiere se llevarán a cabo considerando los tres factores mencionados en los puntos anteriores, además de la variable estrellas, puesto que parece natural plantearse que la cantidad de estrellas asignadas a cada hotel sea un aspecto fundamental a la hora de establecer qué establecimientos forman parte de la competencia de otros.

En este caso se forman los grupos atendiendo, en forma de factores, y con el apoyo de la variable *ESTRELLAS*, a todos los aspectos contemplados en el estudio. Tras llevar a cabo un proceso jerárquico de *cluster,* se concluye que el número óptimo de grupos a crear es 5.

El dendrograma obtenido se presenta a continuación.

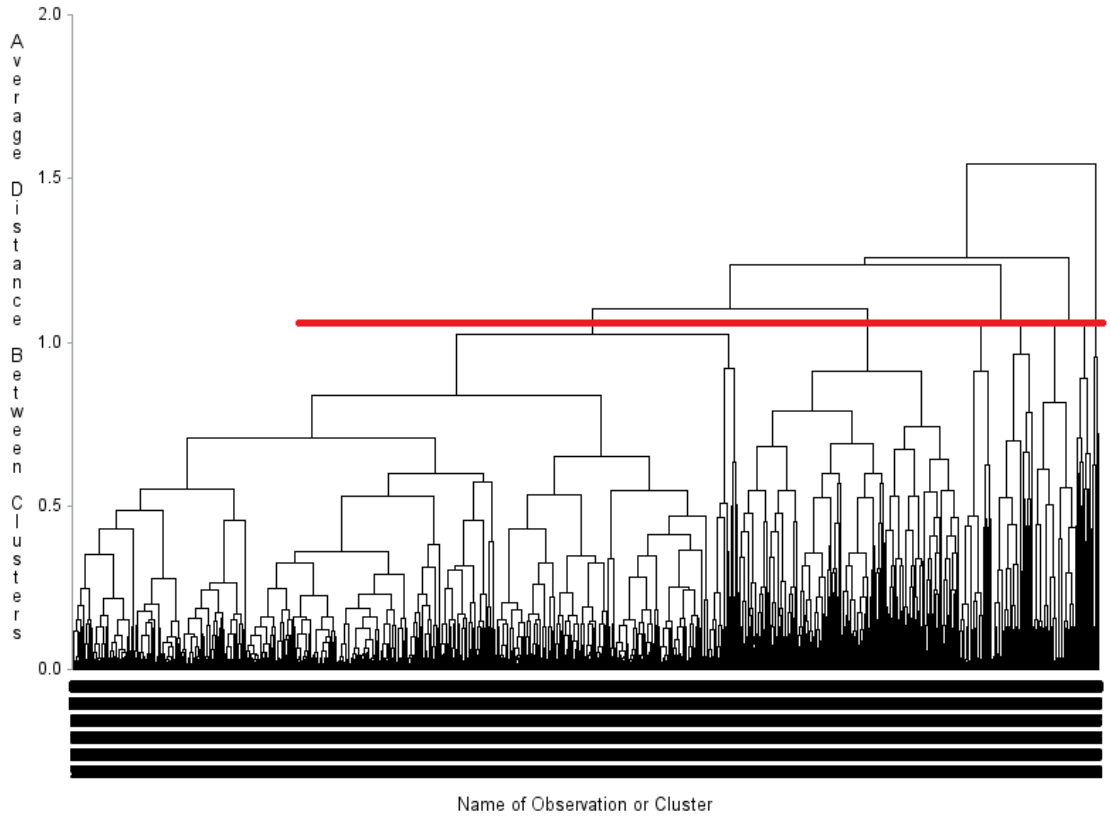

Figura 15. Dendrograma

5.2.2. Factor 1, Factor 2 y Factor 3

A continuación, repetiremos el mismo proceso, pero obviando la intervención de la cantidad de estrellas de cada hotel.En este caso, tras finalizar con el proceso jerárquico, se concluye que el número óptimo de grupos es 8. En la Figura16 se puede ver en dendrograma correspondiente.

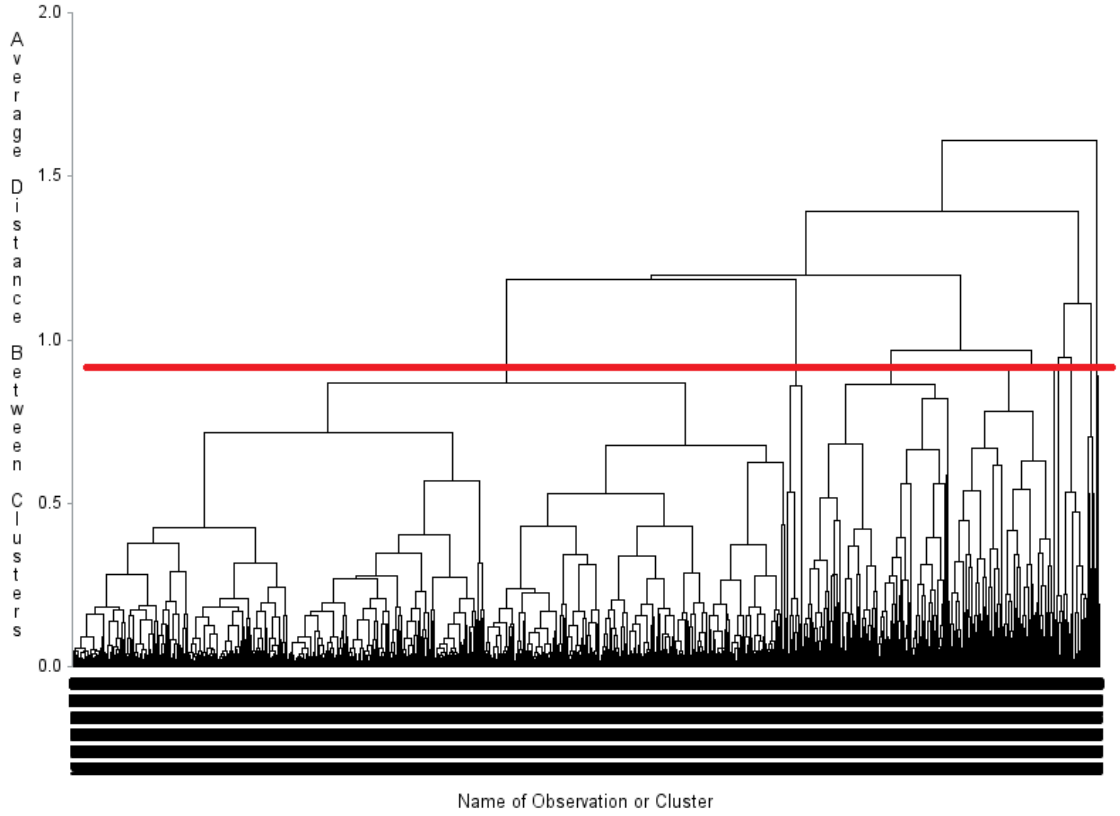

Figura 16. Dendrograma

5.2.3. Factor 1, Factor 2 y Estrellas

Recordando las propiedades de los factores que se están considerando, es importante tener en cuenta que el factor 3 pretendía representar el momento temporal afectado. Parece natural preguntarse cómo funcionará el análisis clúster excluyendo este parámetro, ya que cabe esperar que todos los hoteles tengan un comportamiento similar a la hora de encarecer o disminuir sus precios (con esto no nos referimos a que los distintos hoteles igualen sus tarifas, sino que el encarecimiento o disminución del precio de todos los centros variará de forma parecida atendiendo al momento temporal. Por ejemplo, parece natural pensar, que tanto un hotel de la zona Centro como uno de Chamberí encarecerán sus precios los sábados). Por este motivo, se procede a agrupar los datos excluyendo la información aportada por el factor 3.

En este caso, tras terminar con el proceso de *clúster* jerárquico, se concluye que la cantidad ideal de grupos a crear es 9, uno más que en el caso en que también se consideraba el factor temporal.

En la Figura 17 podemos ver el dendrograma correspondiente a este caso.

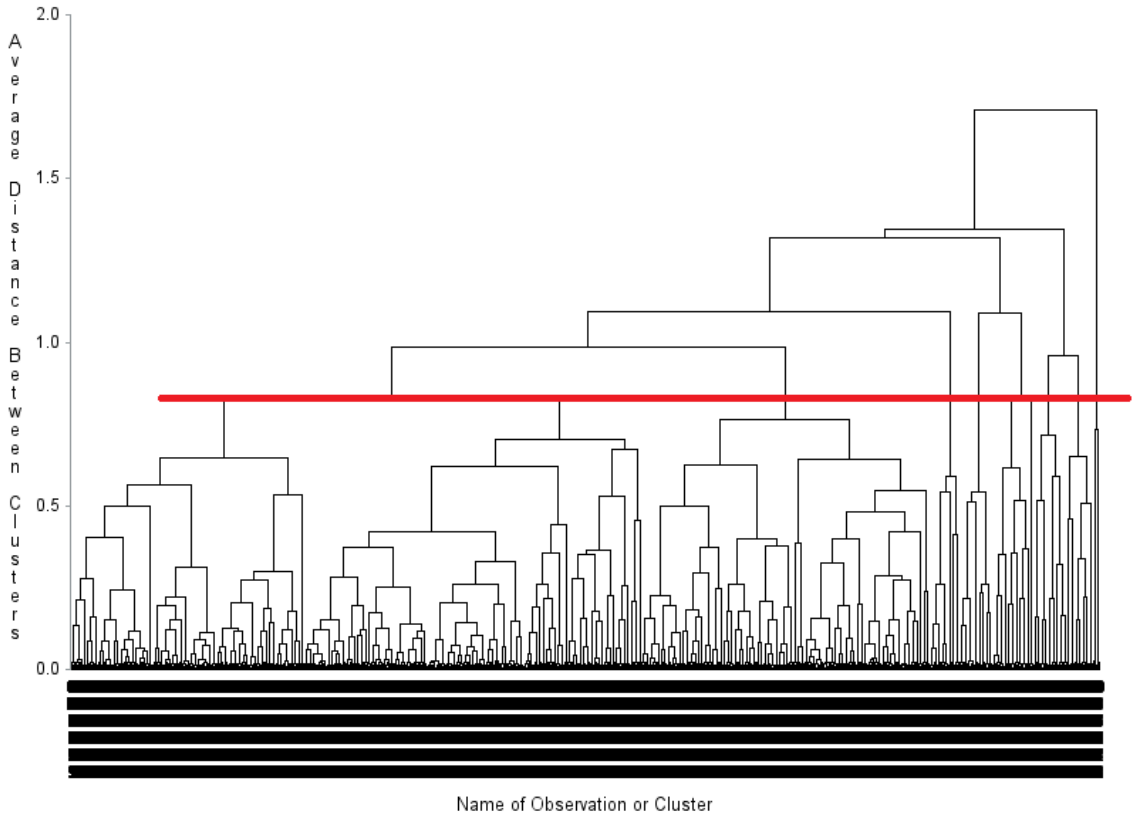

Figura 17. Dendrograma

5.2.4. Factor 1 y Factor 2

En este punto se tratará un caso equivalente al anterior, pero excluyendo la consideración de la cantidad de estrellas de cada establecimiento. En este caso, el número óptimo de grupos a crear es 6.

En el dendrograma recogido en la Figura 18 puede contemplar cómo se distribuyen los datos en este caso.

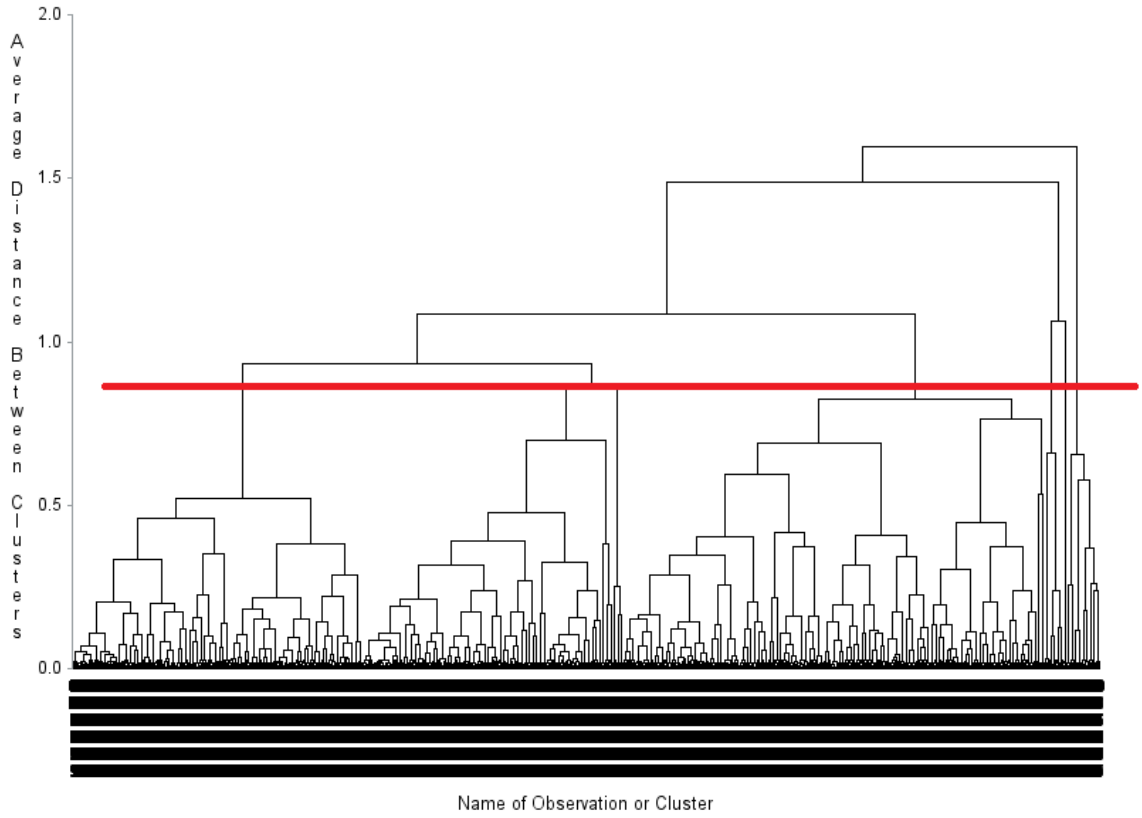

Figura 18

5.3. Estudio de la competencia

A continuación, tomando como base los distintos grupos creados en los procesos de *clúster*, se procede a llevar a cabo un análisis del estudio del comportamiento de los hoteles y su competencia.

Para ello, se tomarán los distintos grupos creados en el análisis *clúster*, entendiendo que para cada hotel, su competencia está formada por los establecimientos que están en su mismo grupo o *clúster*. Nótese que, al haber llevado a cabo distintos procesos de agrupamiento, la competencia de un hotel variará. Se considerarán los distintos casos por separado.

A la hora de establecer si un hotel está dentro de su mercado o no (esto es, si su precio se ajusta al marcado por su competencia), nos fijaremos principalmente en los siguientes datos:

- Precio del hotel a analizar.
- Precio medio de su competencia ($precio_{medio}$).
- Desviación típica del precio de su competencia ($devest$).

Así, consideraremos que un hotel está *dentro de su mercado*, si su precio está dentro del intervalo $(precio_{medio} - 1.96devest, precio_{medio} + 1.96devest)$. En otro caso, se dirá que el hotel en cuestión está fuera de su mercado, esto es, que tiene un precio demasiado caro o barato, teniendo en cuenta el comportamiento de su competencia.

Se analizará el comportamiento de los distintos grupos de competencia y los hoteles que los conforman para cada una de las casuísticas detalladas en el punto anterior.

Con el objetivo de obtener una visión conjunta de los resultados obtenidos de todas ellas, otro de los pasos a seguir consistirá en la ordenación, de mayor a menor precio, de los hoteles que forman cada *clúster*. Así, puesto que la agrupación de los hoteles variará en los distintos casos, tendremos una visión global de la posición general que ocupa un hotel, independientemente de la agrupación que se lleve a cabo, lo cual nos permitirá afirmar fehacientemente si un hotel es, efectivamente muy caro o barato, atendiendo a las diversas variaciones que pueda sufrir la definición de "su competencia directa".

5.3.1. Factor 1, Factor 2, Factor 3 y Estrellas

Recordemos que en este caso, la cantidad de grupos creado es 5, cuyos precios se distribuyen como se puede ver en la Figura 19.

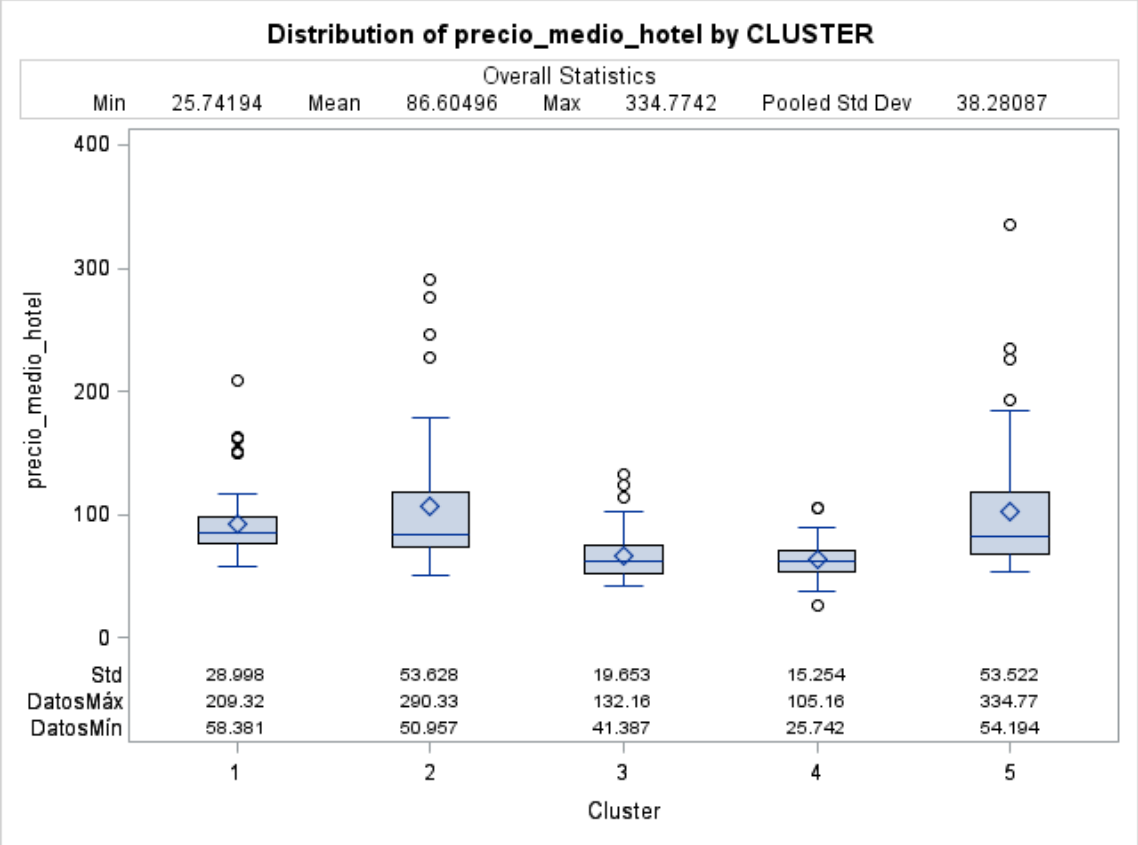

Figura 19. Distribución de los precios por clúster

A simple vista, se puede observar que, en este caso hay menos *outliears* (representados por puntos blancos, cada uno de ellos es un hotel que está "fuera de su mercado"), y los precios están en general más centralizados en sus respectivos grupos. Tomemos como ejemplo el hotel que encabeza nuestra lista, el *Hotel* Europa, con 3 estrellas y situado en la zona centro. En el primer proceso de análisis *clúster*, este hotel cae dentro del primer grupo. Si nos fijamos en este *clúster* (Figura 20), podemos observar que los hoteles que parecen quedarse fuera de su mercado, que está formado por 55 hoteles con precios bastante por encima de su competencia, son:

- *Hotel Urban*(1; 1/55)
- *Hotel Atlántico*(0.9629; 2/55)
- *NH Collection Madrid Palacio de Tepa* (0.9259 ; 3/55)
- *Hotel IberoStar Las Letras* (0.8888; 4/55)
- *ME Madrid Reina Victoria*(0.8518; 5/55).
- *Hotel Catalonia Las Cortes* (0.8148; 6/55).

El primero número de la tupla que acompaña a cada hotel marca la posición de cada hotel en el *clúster* correspondiente, de modo que ordenados de mayor a menor precio (orden indicado por el segundo elemento de la tupla), al hotel más caro se le asigna el número 1 y al hotel más barato el -1. Este tipo de ordenación permitiría llevar a cabo agregaciones de hoteles, un estudio de su distribución de en los distintos *clúster* a los que es asignado y otro tipo de análisis, que permitirían conocer el comportamiento de un establecimiento independientemente de la competencia establecida, proporcionando así una visión más general.

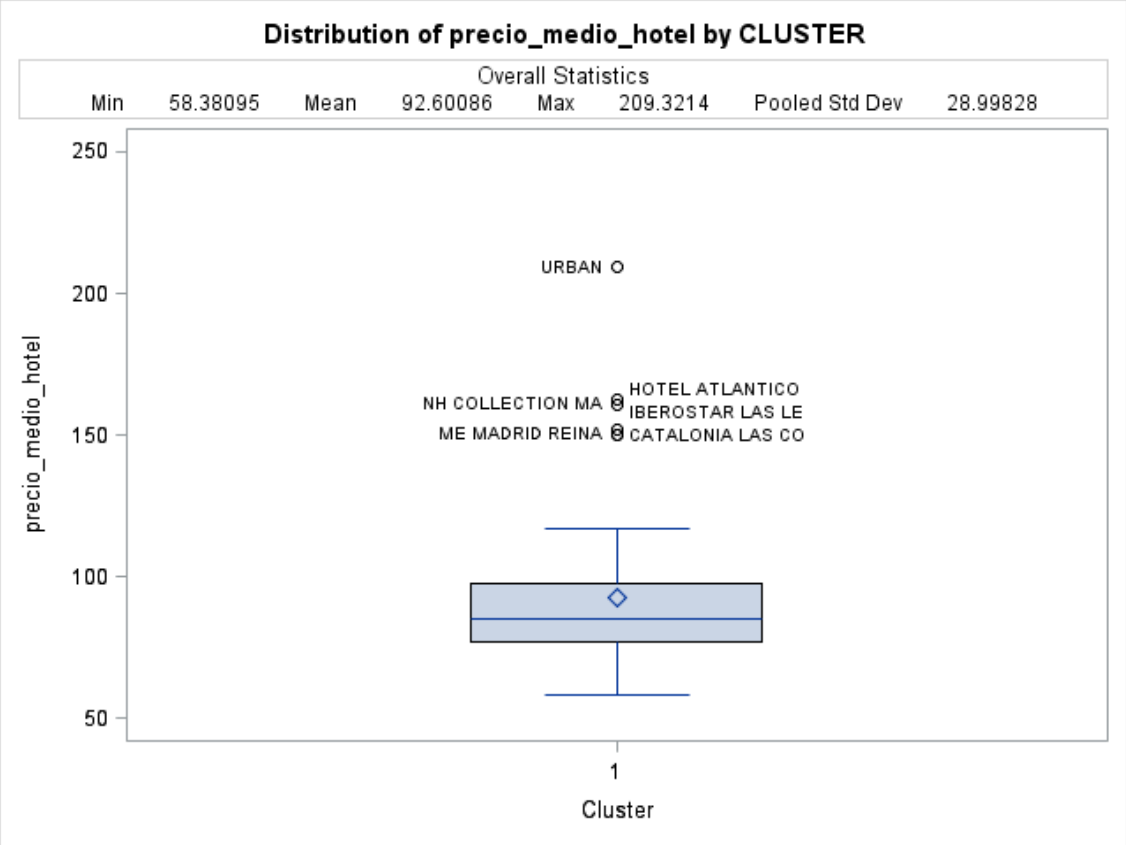

Figura 10. Distribución del precio en el clúster1

Como se puede observar en la Tabla 4, nuestro hotel está dentro de los márgenes marcados, y por tanto, podemos afirmar que está dentro de su mercado.

precio_medio_cluster1	min_cluster1	max_cluster1	DEVEST_c1	HOTEL	precio_medio_hotel	min_h	max_h	Distancia	lim_inf	lim_sup	orden_EUROPA	pos_EUROPA
92.60	58.38	209.32	28.99	HOTEL EUROPA	69,41	64	70	27.69	35.77	149.42	50/55	-0,8148

Tabla 4

Las variables recogidas en esta tabla representan lo siguiente:

- *Precio_medio_c1*: precio promedio del *clúster* 1.
- *Min_c1*: precio mínimo del *clúster* 1.
- *Max_c1*: precio máximo del *clúster* 1.
- *DEVEST_c1*: desviación estándar del *clúster* 1.
- *Hotel*: hotel escogido como ejemplo.
- *Precio_medio_hotel*: precio medio del hotel escogido como ejemplo.
- *Min_h*: precio mínimo del hotel escogido como ejemplo.
- *Max_h*:precio máximo del hotel escogido como ejemplo.
- *Distancia*: diferencia entre el promedio del precio del *clúster* y el del hotel.
- *Lim_inf*: límite inferior tomado para afirmar que un hotel está fuera de su mercado ($precio\_medio\_c - 1.96devest\_c$.
- *Lim_sup*: límite superior tomado para afirmar que un hotel está fuera de su mercado ($precio_{medio_c} + 1.96devest\_c$.
- *Orden_EUROPA*: posición que ocupa el hotel escogido en la lista de hoteles del *clúster* correspondiente ordenada de más caro a más barato.
- *Pos_EUROPA*: posición que ocupa el hotel escogido en su *clúster*.

5.3.2. Factor 1, Factor 2 y Factor 3

En este caso se crean 8 grupos. En el gráfico de cajas y bigotes de la Figura 21, podemos ver cómo se distribuyen los precios en cada uno de los grupos. En general, excepto algunos casos de *outliers*, se puede observar que para cada *clúster* los precios están bastante agrupados.

Tomemos nuevamente como ejemplo el *Hotel* Europa, que vuelve a estar clasificado dentro del primer *clúster*. En este caso, los hoteles que se quedan fuera de este grupo (Figura 22), por establecer unos precios considerablemente más altos que los de su competencia, son:

- *Hotel NH Collection Madrid Palacio de Tepa* (1; 1/32)
- *ME Madrid Reina Victoria*(0.9354; 2/32)
- *Hotel Liabeny*(0.8709; 3/33)

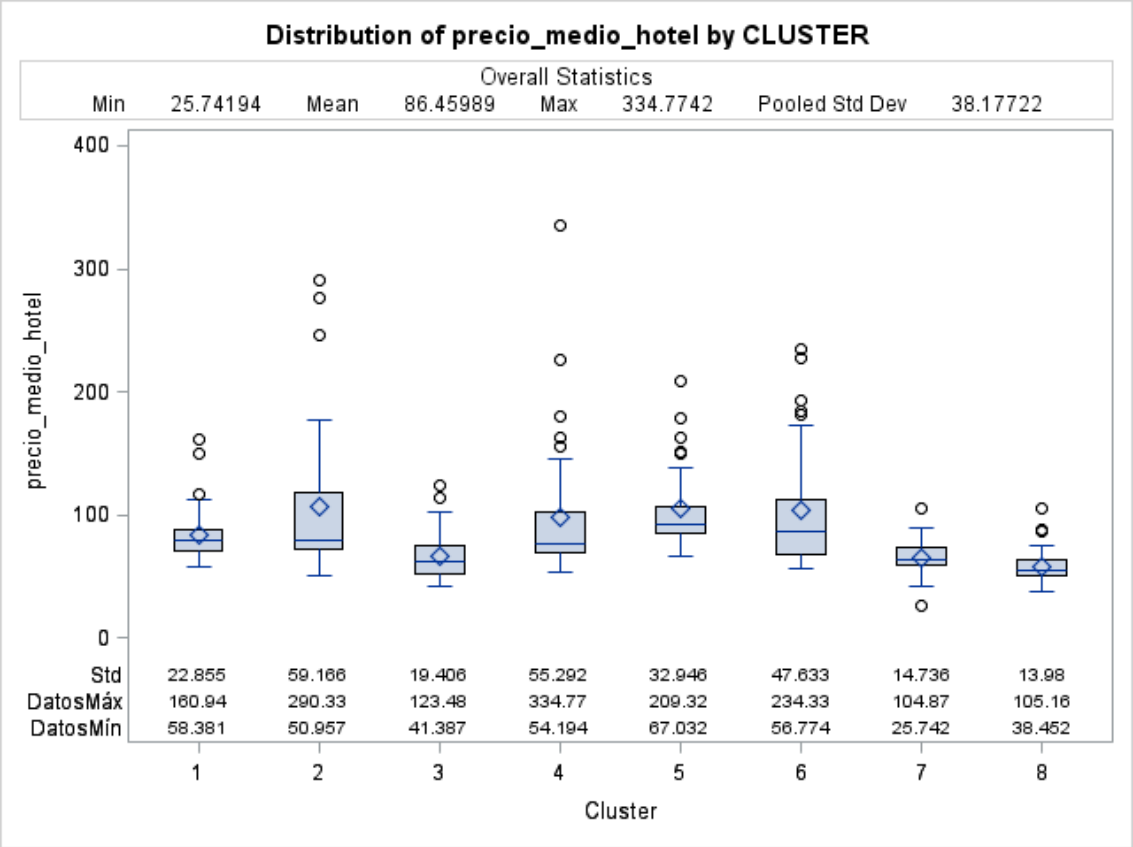

Figura 21. Distribución de los precios por clúster

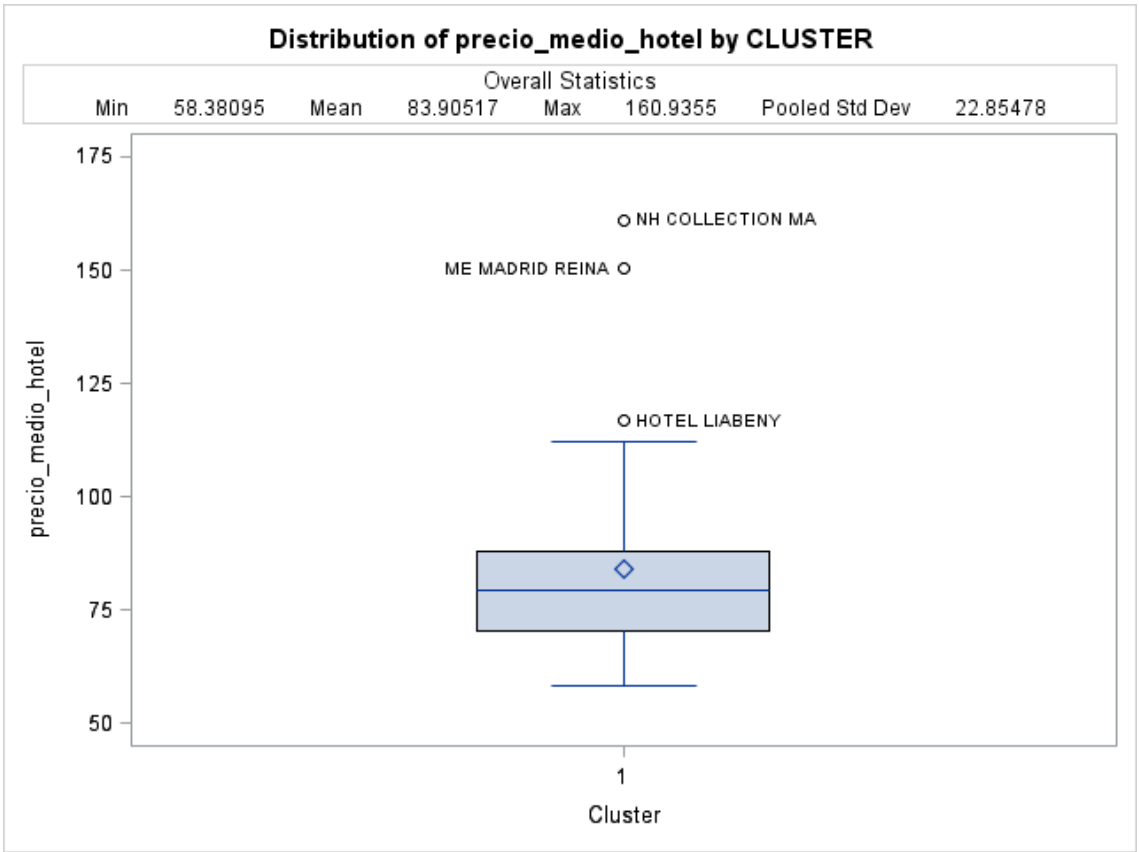

Figura 22. Distribución del precio en el clúster 1.

En la Tabla 5 se puede ver un resumen del comportamiento del clúster afectado y del hotel tomado como ejemplo. Como se puede observar, el precio establecido por el mismo, está dentro de los márgenes marcados, por lo que, en este caso, podemos afirmar que el Hotel Europa está dentro de su mercado. Además, si nos fijamos en la posición que ocupa, aunque está entre las más bajas, su precio no es tan barato como para que ocupe las últimas posiciones del *clúster*.

precio_medio_cluster1	min_cluster1	max_cluster1	DESVT_c1	HOTEL	precio_medio_hotel	min_h	max_h	distancia	lim_inf	lim_sup	orden_EUROPA	pos_EUROPA
83.90	58.38	160.93	22.84	HOTEL EUROPA	69,41	64	70	14.49	39.13	128.66	27/32	-0,6774

Tabla 5

5.3.3. Factor 1, Factor 2 y Estrellas

A continuación desarrollaremos otra vez los pasos expuestos, pero esta vez excluyendo la variable temporal. En el caso que nos ocupa, a la hora de marcar la competencia de cada hotel, intervendrán los factores 1 y 2, además de la variable *ESTRELLAS*.
En este contexto, se crean 9 grupos

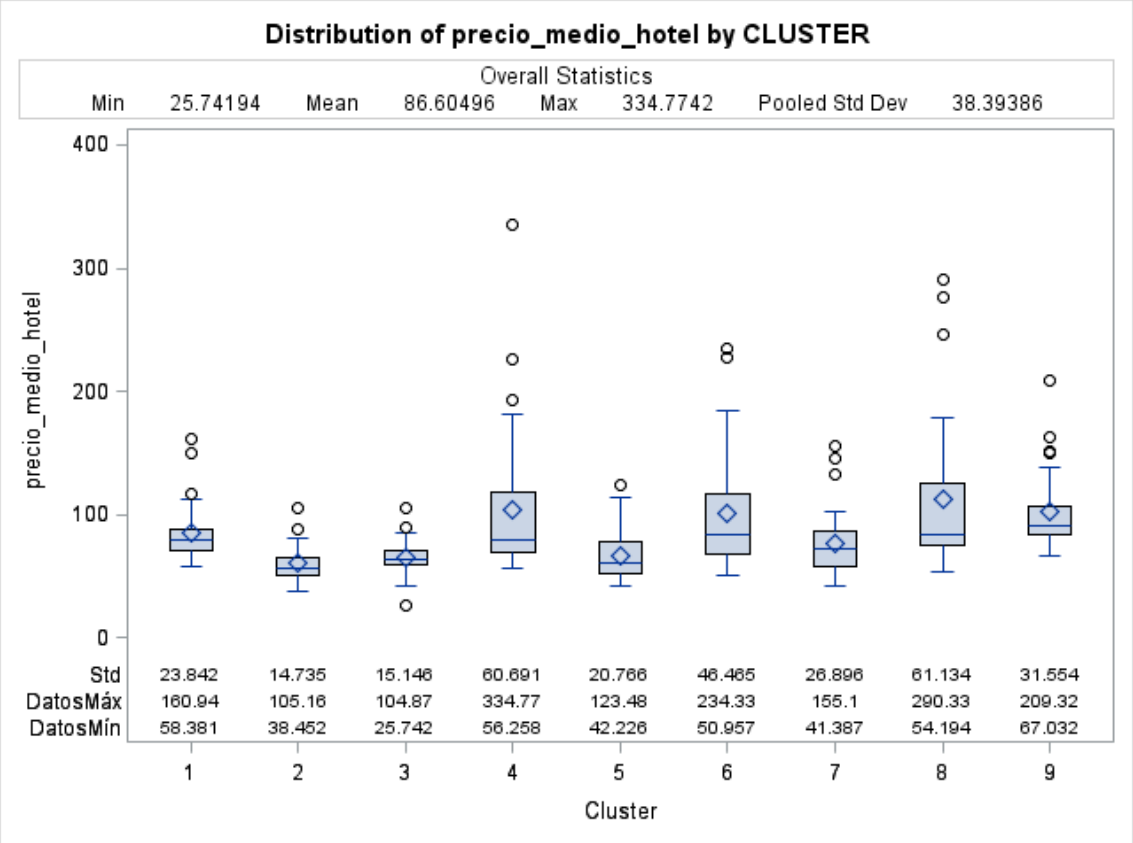

Figura 11. Distribución de los precios por clúster

En este caso, el hotel que estamos tomando como ejemplo está en el grupo número 1, que contiene otros 29 centros, que serán considerados la competencia del *Hotel Europa.*

Observando más detenidamente el primer grupo (Figura 24), podemos ver que tres hoteles se quedan fuera de su mercado, por marcar unos precios que sobresalen con respecto a los establecidos por su competencia. Estos hoteles son los mismos que en el caso anterior (y lógicamente, ocupan la misma posición dentro de su *clúster)*:

- *Hotel NH Collection Madrid Palacio de Tepa* (1; 1/29)
- *ME Madrid Reina Victoria*(0.9258; 2/29)
- *Hotel Liabeny*(0.8571; 3/29)

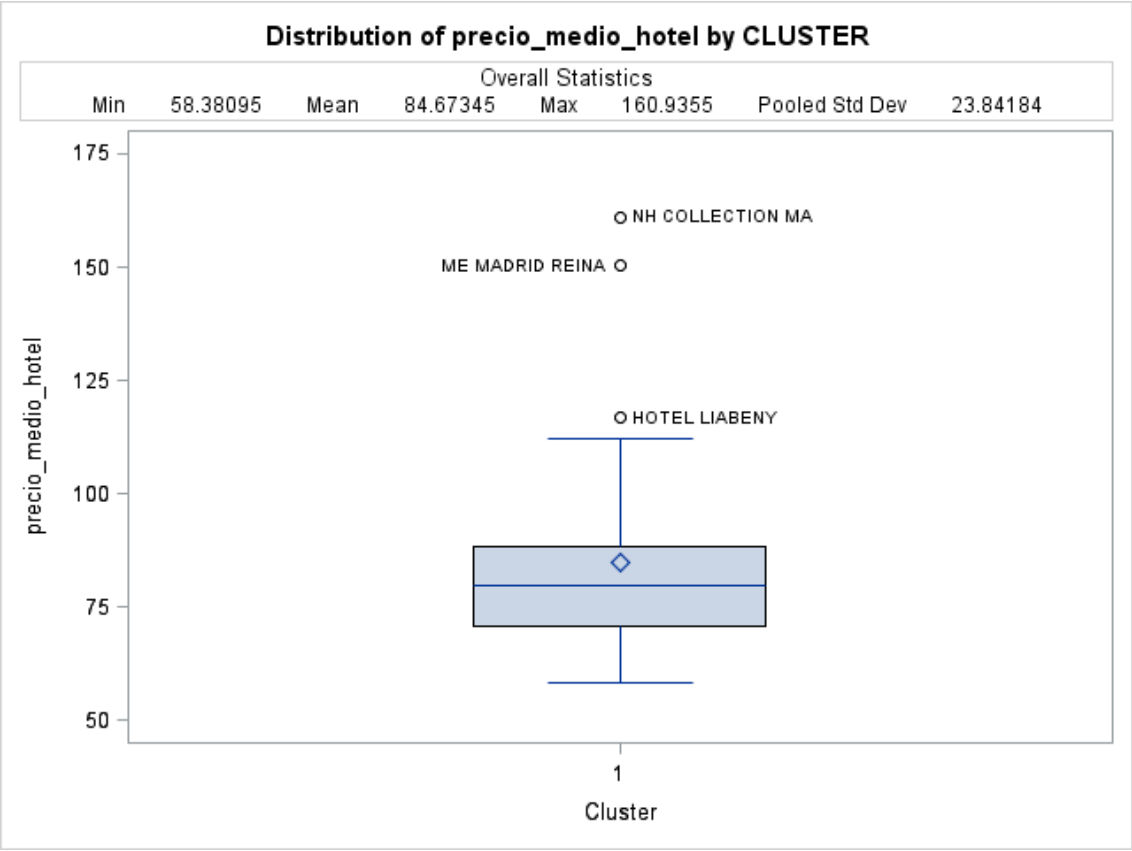

Figura 24. Distribución del precio en el primer clúster

Esto nos indica que, para esta definición de competencia, el *Hotel Europa* está dentro de su mercado. En Tabla 6 se pueden encontrar más detalles sobre ello.

precio_medio_c1	min_c1	max_c1	DEVEST_c1	HOTEL	precio_medio_hotel	min_h	max_h	distancia	lim_inf	lim_sup	orden_EUROPA	pos_EUROPA
84,67	40	160.93	23/84	HOTEL EUROPA	69,41	64	70	15,26	37.94	131.39	24/29	-0,6428

Tabla 6

5.3.4. Factor 1 y Factor 2

Como último caso se plantea una situación análoga a la anterior, aunque esta vez, además del factor temporal, se excluirá la variable *ESTRELLA*, de modo que sólo intervendrán los factores 1 y 2.Así, para establecer los diferentes grupos se tendrán en cuenta tanto las valoraciones recibidas por los hoteles, como su ubicación en la ciudad. En este contexto, el número de *clústers* a crear es 6, cuyos comportamientos en cuanto a establecer un precio se refiere, quedan reflejados en el siguiente gráfico de cajas y bigotes (Figura 25).

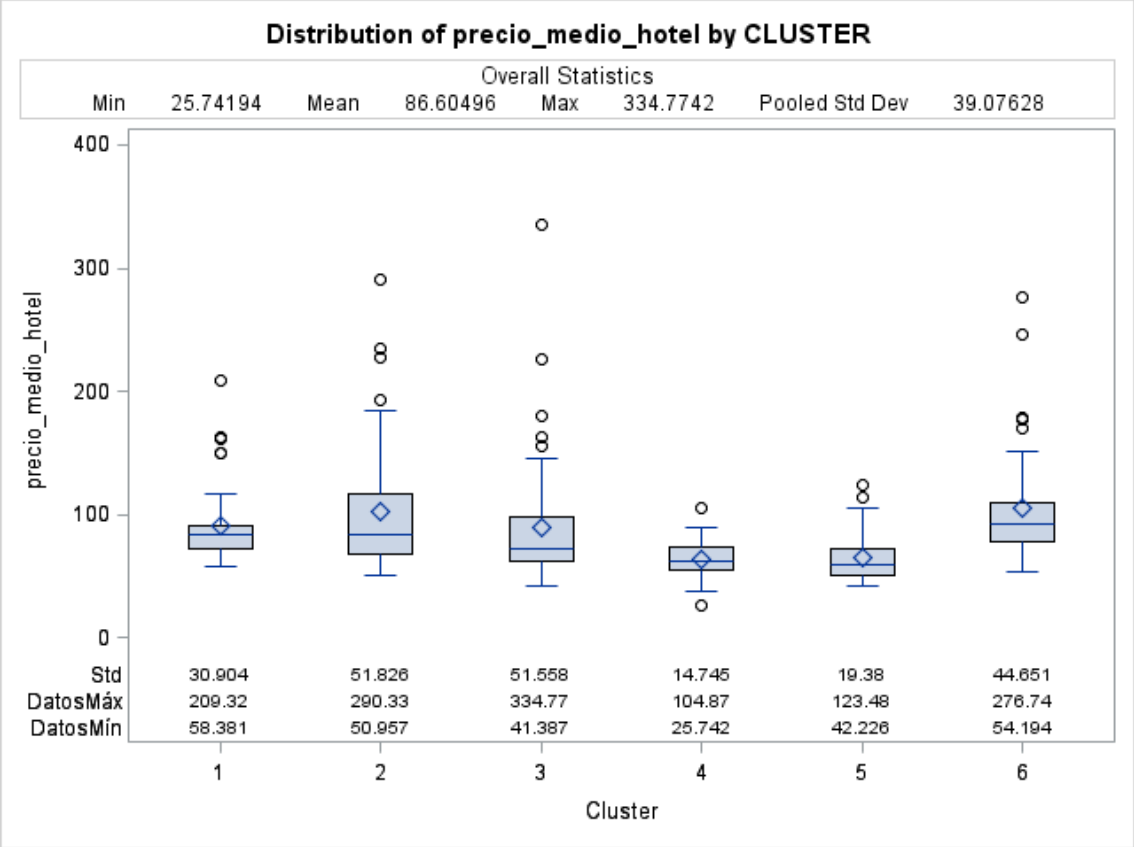

Figura 25. Distribución de los precios por clúster

Analizamos con más detalle el grupo número 1 (Figura 26), al cual pertenece el *Hotel Europa.* En este grupo hay un total de 44 hoteles, de los cuales tres quedan fuera de su mercado, a saber:

- *Hotel Urban*(1; 1/44)
- *Hotel Atlántico*(0.9534; 2/44)
- *NH Collection Madrid Palacio de Tepa* (0.9069; 3/44)
- *ME Madrid Reina Victoria* (0.8604; 4/44).
- *Hotel Catalonia Las Cortes* (0.8139; 5/44)

Nuevamente, como se puede observar en la Tabla 7, el hotel que nos ocupa está dentro de su mercado, a la hora de comparar sus precios con los establecidos por su competencia.

precio_medio_c1	min_c1	max_c1	DESVEST_c1	HOTEL	precio_medio_hotel	min_h	max_h	distancia	lim_inf	lim_sup	orden_EUROPA	pos_EUROPA
90.93	58.38	208,32	30.90	HOTEL EUROPA	69,41	64	70	26.02	30.36	151.49	39/44	-0,7674

Tabla 7

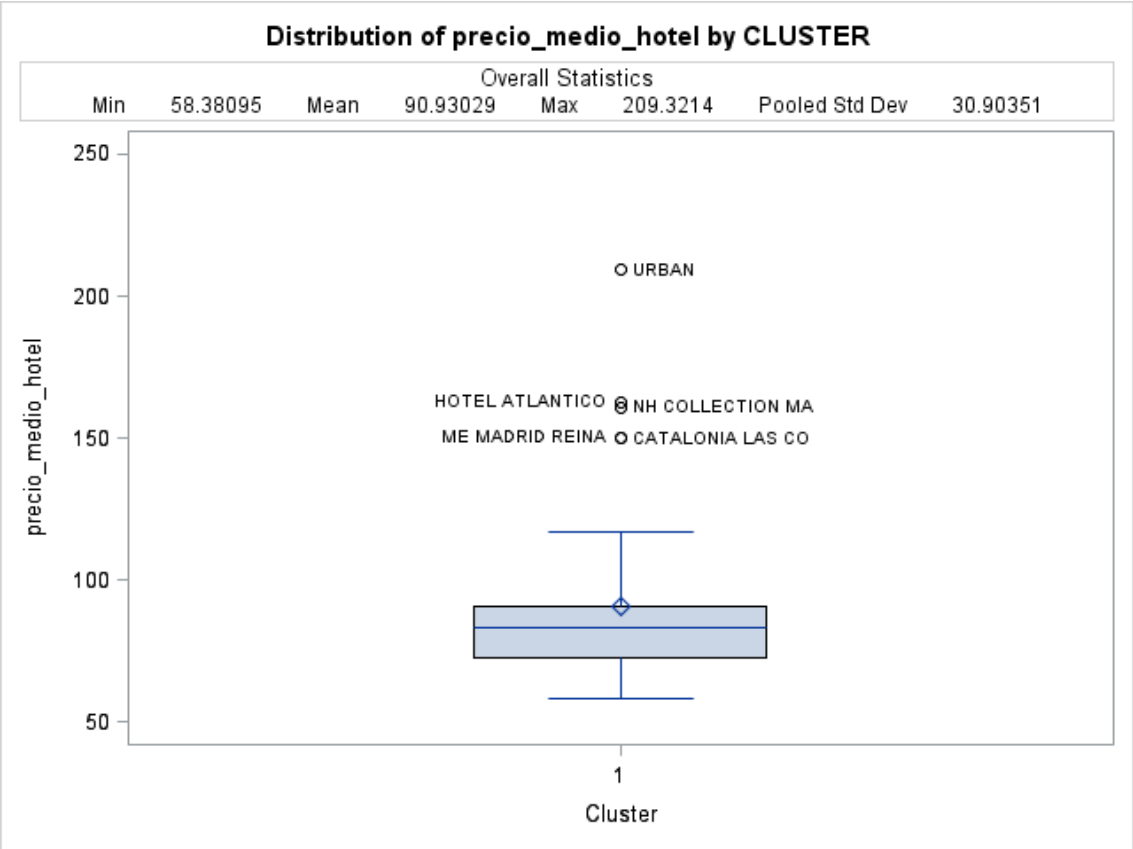

Figura 26. Distribución de los precios en el primer clúster.

5.4. Conclusión

Como se ha detallado en las páginas anteriores, para las distintas definiciones de competencia establecidas, estructuradas atendiendo a distintas características de los hoteles, la empresa en que se ha focalizado la atención siempre está dentro de su mercado. Podemos interpretar esta afirmación como que los precios marcados por el *Hotel Europa* están bastante adecuados con respecto a la competencia directa que le podría suponer una reducción en el número de clientes. Además de que su precio se encuentra dentro de los márgenes marcados, su posición en el ránking de su *clúster*, aunque baja, nunca toma las últimas posiciones. Esta situación se repite para las distintas agrupaciones llevadas a cabo, en que el hotel mencionado siempre ocupa una

de las 6 últimas posiciones de su grupo. Esto nos da una idea más precisa del comportamiento de este establecimiento, ya que, aún variando ligeramente los hoteles con que se le compara, siempre ocupa uno de los puestos más bajos, aún estando dentro de su mercado, lo cual nos podría etiquetar este hotel de "barato", de acuerdo a sus características.

En este caso, hemos escogido un hotel y el *clúster* que se le asigna para ejemplificar el proceso, puesto que los pasos a seguir son idénticos para otros supuestos.
Las conclusiones que podemos sacar analizando los resultado obtenidos es, que, si nos fijamos en las cuatro situaciones estudiadas, en todas ellas los hoteles *ME Madrid Reina Victoria* y *NH Collection Madrid Palacio de Tepa* están fuera de su mercado, ocupando las primeras posiciones en el ránking de su *clúster*. Esto nos permite afirmar que estos hoteles establecen unos precios demasiado altos, puesto que, independientemente de la estructura de competencia observada, siempre quedan por encima de su mercado.

Para un hotel es de vital importancia ajustar sus precios a los marcados por otros establecimientos con que puedan ser comparados. Tan importante es no quedarse fuera del mercado a la baja, como le ocurre al hotel *Aravaca Garden* en la última situación expuesta, representado por el *outlier* que queda por debajo del grupo 3, como no hacerlo por subir demasiado los precios, como ocurre por ejemplo, con el mencionado *ME Madrid Reina Victoria*¸o con el *Hotel VillaMagna*, representado por el punto más alto que queda fuera del tercer *clúster*. En la primera situación, el problema con que se enfrentaría el hotel en cuestión es que no está ganando tanto dinero como podría, asumiendo que un cliente cualquiera, estaría dispuesto a pagar por el aproximadamente el mismo dinero que por cualquier otro de los establecimientos que forman parte de su grupo o competencia. En cuanto a la segunda situación, el problema para el hotel podría agravarse, ya que en este caso no es tanto ganar "poco" dinero, como perder clientes que, en situaciones normales, se decantarán por las opciones más baratas dentro de un abanico de hoteles considerados "equivalentes" (esta equivalencia de cara a los clientes es lo que se ha venido entendiendo como competencia).

6. Posibilidades para el futuro

A continuación se detallan ciertos aspectos que podrían suponer una continuación de este proyecto y que, lamentablemente no se han podido desarrollar por razones de tiempo.

- Análisis en profundidad del estudio predictivo. Para ello podría ser una buena idea contar con los datos de cada hotel recogidos para más de un mes. Por ejemplo, conociendo los precios de todos los establecimientos para todo un año, se podría estudiar cómo afectan los distintos meses a la fluctuación de los mismos. Conocer la demanda de cada establecimiento para cada día, o la cantidad de habitaciones disponibles que ofrece son otras medidas que aportarían una información útil a la hora de predecir el precio de los distintos hoteles.

  Además, se podría profundizar en la aplicación de las técnicas utilizadas, buscando nuevas variantes entre sus parámetros, o aplicar otras diferentes, como pueda ser *Gradient Boosting*, que es otro algoritmo típico relativo a árboles de decisión.

- Análisis más profundo de la competencia hotelera. Siguiente la línea de trabajo expuesta en el anterior epígrafe y profundizando en más detalles, se podría generalizar el proceso seguido. Para ello, habría que ordenar los hoteles de los distintos *clústers* para todas las situaciones, de modo que siempre podríamos saber la posición que ocupa en un hotel en su grupo, independientemente de la definición de competencia seleccionada. Esto nos permitiría observar qué hoteles ocupan siempre las posiciones más altas o más bajas de su grupo, estudiando la variación de la posición de un hotel en su grupo atendiendo al conjunto de establecimientos que se consideren su competencia. No es más que repetir los pasos detallados, haciéndolo para todos los hoteles y *clústers* obtenidos, medida que nos permitirá analizar hotel por hotel su comportamiento, además de ver hasta qué punto su competencia le afecta, en el sentido de que su precio pueda resultar fuera de límite o por el contrario, un precio "aceptable", dependiendo de con qué se le compare.

- Análisis de la variación del comportamiento de un hotel atendiendo a su competencia. Para ello, además de tener clara la competencia a tener en cuenta para el establecimiento seleccionado, es necesario conocer su demanda, y tener una función que la represente. Así, se podría calcular el equilibrio de Nash del hotel escogido para después fijar su función de mejor respuesta, que proporcionará el precio óptimo que debería fijar este hotel, atendiendo a la variación del comportamiento de su competencia, de modo que este precio establecido no sea tan caro como para que un cliente escoja un centro de la competencia, ni tan barato como para que a este hotel le siga quedando margen de "ganar más".

7. Anexos

7.1. Anexos Descriptivos

7.1.1. Variables Cualitativas

Variables Cualitativas				
Nombre de la Variable	Tipo	Nº Niveles	Detalles	moda
Nombre Hotel	Categórica	278	nombre del hotel correspondiente	-
H	Categórica	278	identificador numérico de los hoteles (ordenador por su distancia al centro de la ciudad)	-
ESTRELLAS	Ordinal	5	estrellas asignadas a cada hotel, variable importante a la hora de definir la *competencia*	4
ZONA	Categórica	16	distrito financiero en que se encuentra el establecimiento, variable importante a la hora de definir la *competencia*	centro
FECHA	Ordinal	31	Día del mes	-
FECHAINSC	Ordinal	253	fecha de registro del hotel en *Booking*	18/01/2010
CODPOSTAL	Ordinal	39	código postal del establecimiento, nos permitirá definir la *competencia* de una forma mucho más ajustada de lo permitido por el distrito financiero	28013
USABLE	Dicotómica	2	nos dirá si un hotel es "usable" en algunos de los puntos del estudio, en el sentido de que conocemos su precio para todas las noches del mes (lo cual nos informa, además, de que este hotel tenía al menos una habitación disponible en el momento de la extracción de datos)	1
ESTACIONAL	Dicotómica	2	nos permite afirmar si un hotel aumenta o no sus precios los fines de semana	1
OFERTA	Dicotómica	2	nos da información sobre si los distintos precios ofrecidos en la noche i están sujetos a ofertas especiales o no	-
DIA	Categórica	7	día de la semana	-
FINDE	Categórica	3	Categoriza las observaciones, atendiendo a si el día siguiente al correspondiente es laboral; el día en cuestión es laboral y el siguiente no, o bien el día en cuestión y el siguiente no son laborales.	-

Tabla 8

7.1.2. Variables Cuantitativas

Variables Cuantitativas						
Nombre de la Variable	Tipo	Detalles	Mínimo	Mediana	Máximo	Media
PRECIO	Continua	precio de los distintos hoteles en la noche i-ésima	24	72	405	85.95
PROMEDIO	Continua	promedio del precio de cada hotel	25,84	74,52	334,77	86,41
COMENTARIOS	Discreta	nº de comentarios recibidos por cada establecimiento, servirá un medidor de la "popularidad" de cada hotel	2	1014	7467	1524
CALIDADPRECIO	Continua	valoración sobre la calidad/precio	5,5	7,9	9,4	7,9
CONFORT	Continua	valoración sobre el confort ofrecido por el establecimiento	5,8	8,2	9,8	8,14
INSTALACIONESSERVICIOS	Continua	valoración sobre las instalaciones y los servicios del hotel	6	8	9,7	7,94
LIMPIEZA	Continua	valoración sobre la limpieza	6,2	8,6	10	8,52
PERSONAL	Continua	valoración sobre el personal de servicio del hotel	0,6	8,5	9,9	8,44
UBICACIÓN	Continua	valoración sobre la ubicación del establecimiento	6	8,8	9,8	8,58
VALGLOBAL	Continua	valoración global sobre le hotel	0	8,3	9,4	8,21
WIFI	Continua	valoración sobre el servicio WiFi ofrecido	4,1	8,1	10	7,9

Tabla 9

7.1.3. Inferencia Estadística

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-88,287184	7,420101	-11,9
CALIDADPRECIO	1	-48,705748	1,249825	-38,97
CONFORT	1	16,765555	1,823883	9,19
INSTALACIONESSERVICI	1	36,82388	2,317682	15,89
PERSONAL	1	10,269009	1,275428	8,05
UBICACION	1	14,663264	0,530038	27,66
VALGLOBAL	1	-12,35586	0,631155	-19,58
WIFI	1	2,396863	0,580924	4,13

Tabla 10

Comparisons significant at the 0.05 level are indicated by ***.				
ESTRELLAS	Difference between means	95% Confidence Limits		
5 4	975.252	949.245	1.001.259	***
5 3	1.099.459	1.071.495	1.127.424	***
5 2	1.201.568	1.166.052	1.237.084	***
5 1	1.360.628	1.306.139	1.415.116	***
4 5	-975.252	-1.001.259	-949.245	***
4 3	124.207	107.278	141.135	***
4 2	226.316	198.640	253.991	***
4 1	385.375	335.640	435.110	***
3 5	-1.099.459	-1.127.424	-1.071.495	***
3 4	-124.207	-141.135	-107.278	***
3 2	102.109	72.586	131.632	***
3 1	261.168	210.382	311.954	***
2 5	-1.201.568	-1.237.084	-1.166.052	***
2 4	-226.316	-253.991	-198.640	***
2 3	-102.109	-131.632	-72.586	***
2 1	159.059	103.755	214.364	***
1 5	-1.360.628	-1.415.116	-1.306.139	***
1 4	-385.375	-435.110	-335.640	***
1 3	-261.168	-311.954	-210.382	***
1 2	-159.059	-214.364	-103.755	***

Tabla 11

COEFICIENTES DE CORRELACIÓN DE PEARSON	
FECHAINSC	0.04096
	0.0002
	8382
OFERTA	-0.12485
	<.0001
	8382
FECHA	0.02819
	0.0098
	8382
ESTRELLAS	0.49625
	<.0001
	8382
CODPOSTAL	-0.20624
	<.0001
	8382
COMENTARIOS	-0.18114
	<.0001
	8382
FINDE	0.08107
	<.0001
	8382
ZONA	-0.09260
	<.0001
	8382
DIA	0.03348
	0.0022
	8382
VALGLOBAL	0.16399
	<.0001
	8382
LIMPIEZA	0.42149
	<.0001
	8382
CONFORT	0.42722
	<.0001
	8382
UBICACION	0.44199

	<.0001
	8382
INSTALACIONESSERVICIOS	0.45335
	<.0001
	8382
PERSONAL	0.44421
	<.0001
	8382
CALIDADPRECIO	0.11233
	<.0001
	8382
WIFI	0.29001
	<.0001
	8382

Tabla 12

Comparisons significant at the 0.05 level are indicated by ***.				
DIA COMPARASION	Difference between means	95% Confidence Limits		
S - V	4.801	0.956	8.647	***
S - M	9.754	6.130	13.379	***
S - X	10.177	6.555	13.799	***
S - J	10.512	6.701	14.322	***
S - L	10.701	7.076	14.325	***
S - D	10.953	7.133	14.774	***
V - S	-4.801	-8.647	-0.956	***
V - M	4.953	1.349	8.557	***
V - X	5.376	1.775	8.977	***
V - J	5.711	1.919	9.502	***
V - L	5.900	2.296	9.504	***
V - D	6.152	2.351	9.954	***
M - S	-9.754	-13.379	-6.130	***
M - V	-4.953	-8.557	-1.349	***
M - X	0.423	-2.941	3.787	
M - J	0.757	-2.809	4.324	
M - L	0.947	-2.421	4.314	
M - D	1.199	-2.378	4.777	
X - S	-10.177	-13.799	-6.555	***
X - V	-5.376	-8.977	-1.775	***
X - M	-0.423	-3.787	2.941	
X - J	0.334	-3.230	3.898	
X - L	0.524	-2.841	3.888	
X - D	0.776	-2.799	4.351	
J - S	-10.512	-14.322	-6.701	***
J - V	-5.711	-9.502	-1.919	***
J - M	-0.757	-4.324	2.809	
J - X	-0.334	-3.898	3.230	
J - L	0.189	-3.378	3.756	
J - D	0.442	-3.324	4.208	
L - S	-10.701	-14.325	-7.076	***
L - V	-5.900	-9.504	-2.296	***
L - M	-0.947	-4.314	2.421	
L - X	-0.524	-3.888	2.841	
L - J	-0.189	-3.756	3.378	
L - D	0.253	-3.325	3.831	
D - S	-10.953	-14.774	-7.133	***
D - V	-6.152	-9.954	-2.351	***
D - M	-1.199	-4.777	2.378	
D - X	-0.776	-4.351	2.799	
D - J	-0.442	-4.208	3.324	
D - L	-0.253	-3.831	3.325	

Tabla 13

Comparisons significant at the 0.05 level are indicated by ***.				
ZONA	Difference between means	95% Confidence Limits		
RETIRO – CARABA	18.045	6.559	29.531	***
RETIRO – SALAMA	19.138	13.618	24.659	***
RETIRO – CHAMBE	22.495	16.303	28.687	***
RETIRO – CENTRO	25.226	20.071	30.381	***
RETIRO – CHAMAR	46.514	39.728	53.300	***
RETIRO – BARAJA	48.957	42.720	55.195	***
RETIRO – ARGANZ	49.711	42.282	57.139	***
RETIRO – FUENCA	49.798	42.625	56.971	***
RETIRO – MONCLO	53.469	47.224	59.714	***
RETIRO – TETUAN	55.622	49.294	61.951	***
RETIRO - SAN BL	60.281	54.283	66.279	***
RETIRO – CIUDAD	62.583	55.787	69.379	***
RETIRO – HORTAL	64.854	56.641	73.068	***
RETIRO – USERA	75.706	60.235	91.177	***
RETIRO – VILLA	77.771	66.285	89.257	***
CARABA – RETIRO	-18.045	-29.531	-6.559	***
CARABA – SALAMA	1.094	-9.556	11.744	
CARABA – CHAMBE	4.451	-6.562	15.464	
CARABA – CENTRO	7.181	-3.284	17.646	
CARABA – CHAMAR	28.469	17.112	39.827	***
CARABA – BARAJA	30.913	19.874	41.952	***
CARABA – ARGANZ	31.666	19.913	43.419	***
CARABA – FUENCA	31.753	20.160	43.347	***
CARABA – MONCLO	35.424	24.381	46.468	***
CARABA – TETUAN	37.578	26.487	48.668	***
CARABA - SAN BL	42.236	31.331	53.141	***
CARABA – CIUDAD	44.538	33.174	55.902	***
CARABA – HORTAL	46.810	34.545	59.074	***
CARABA – USERA	57.661	39.708	75.614	***
CARABA – VILLA	59.726	45.067	74.384	***
SALAMA – RETIRO	-19.138	-24.659	-13.618	***
SALAMA – CARABA	-1.094	-11.744	9.556	
SALAMA – CHAMBE	3.357	-1.096	7.810	
SALAMA – CENTRO	6.088	3.247	8.928	***
SALAMA – CHAMAR	27.376	22.128	32.624	***
SALAMA – BARAJA	29.819	25.302	34.336	***
SALAMA – ARGANZ	30.572	24.516	36.629	***
SALAMA – FUENCA	30.660	24.919	36.400	***
SALAMA – MONCLO	34.331	29.803	38.858	***
SALAMA – TETUAN	36.484	31.842	41.126	***
SALAMA - SAN BL	41.143	36.963	45.322	***
SALAMA – CIUDAD	43.444	38.183	48.706	***
SALAMA – HORTAL	45.716	38.719	52.713	***
SALAMA – USERA	56.568	41.706	71.429	***
SALAMA – VILLA	58.632	47.982	69.282	***
CHAMBE – RETIRO	-22.495	-28.687	-16.303	***
CHAMBE – CARABA	-4.451	-15.464	6.562	
CHAMBE – SALAMA	-3.357	-7.810	1.096	
CHAMBE – CENTRO	2.731	-1.261	6.722	
CHAMBE – CHAMAR	24.019	18.069	29.969	***
CHAMBE – BARAJA	26.462	21.145	31.779	***
CHAMBE – ARGANZ	27.215	20.541	33.890	***
CHAMBE – FUENCA	27.303	20.914	33.691	***
CHAMBE – MONCLO	30.974	25.648	36.299	***
CHAMBE – TETUAN	33.127	27.704	38.551	***
CHAMBE - SAN BL	37.786	32.752	42.819	***
CHAMBE – CIUDAD	40.087	34.125	46.049	***
CHAMBE – HORTAL	42.359	34.821	49.897	***
CHAMBE – USERA	53.211	38.087	68.334	***

CHAMBE – VILLA	55.275	44.262	66.288	***
CENTRO – RETIRO	-25.226	-30.381	-20.071	***
CENTRO – CARABA	-7.181	-17.646	3.284	
CENTRO – SALAMA	-6.088	-8.928	-3.247	***
CENTRO – CHAMBE	-2.731	-6.722	1.261	
CENTRO – CHAMAR	21.288	16.426	26.150	***
CENTRO – BARAJA	23.731	19.669	27.794	***
CENTRO – ARGANZ	24.485	18.759	30.210	***
CENTRO – FUENCA	24.572	19.182	29.962	***
CENTRO – MONCLO	28.243	24.169	32.317	***
CENTRO – TETUAN	30.397	26.196	34.597	***
CENTRO - SAN BL	35.055	31.371	38.738	***
CENTRO – CIUDAD	37.357	32.480	42.233	***
CENTRO – HORTAL	39.628	32.916	46.341	***
CENTRO – USERA	50.480	35.751	65.210	***
CENTRO – VILLA	52.545	42.079	63.010	***
CHAMAR – RETIRO	-46.514	-53.300	-39.728	***
CHAMAR – CARABA	-28.469	-39.827	-17.112	***
CHAMAR – SALAMA	-27.376	-32.624	-22.128	***
CHAMAR – CHAMBE	-24.019	-29.969	-18.069	***
CHAMAR – CENTRO	-21.288	-26.150	-16.426	***
CHAMAR – BARAJA	2.443	-3.555	8.441	
CHAMAR – ARGANZ	3.196	-4.032	10.425	
CHAMAR – FUENCA	3.284	-3.682	10.250	
CHAMAR – MONCLO	6.955	0.949	12.961	***
CHAMAR – TETUAN	9.108	3.016	15.201	***
CHAMAR - SAN BL	13.767	8.018	19.515	***
CHAMAR – CIUDAD	16.069	9.492	22.645	***
CHAMAR – HORTAL	18.340	10.307	26.373	***
CHAMAR – USERA	29.192	13.815	44.568	***
CHAMAR – VILLA	31.256	19.899	42.614	***
BARAJA – RETIRO	-48.957	-55.195	-42.720	***
BARAJA – CARABA	-30.913	-41.952	-19.874	***
BARAJA – SALAMA	-29.819	-34.336	-25.302	***
BARAJA – CHAMBE	-26.462	-31.779	-21.145	***
BARAJA – CENTRO	-23.731	-27.794	-19.669	***
BARAJA – CHAMAR	-2.443	-8.441	3.555	
BARAJA – ARGANZ	0.753	-5.964	7.470	
BARAJA – FUENCA	0.841	-5.593	7.274	
BARAJA – MONCLO	4.512	-0.867	9.890	
BARAJA – TETUAN	6.665	1.190	12.141	***
BARAJA - SAN BL	11.323	6.234	16.413	***
BARAJA – CIUDAD	13.625	7.616	19.635	***
BARAJA – HORTAL	15.897	8.321	23.473	***
BARAJA – USERA	26.749	11.606	41.891	***
BARAJA – VILLA	28.813	17.774	39.852	***
ARGANZ – RETIRO	-49.711	-57.139	-42.282	***
ARGANZ – CARABA	-31.666	-43.419	-19.913	***
ARGANZ – SALAMA	-30.572	-36.629	-24.516	***
ARGANZ – CHAMBE	-27.215	-33.890	-20.541	***
ARGANZ – CENTRO	-24.485	-30.210	-18.759	***
ARGANZ – CHAMAR	-3.196	-10.425	4.032	
ARGANZ – BARAJA	-0.753	-7.470	5.964	
ARGANZ – FUENCA	0.087	-7.506	7.681	
ARGANZ – MONCLO	3.758	-2.965	10.482	
ARGANZ – TETUAN	5.912	-0.890	12.713	
ARGANZ - SAN BL	10.570	4.075	17.065	***
ARGANZ – CIUDAD	12.872	5.634	20.110	***
ARGANZ – HORTAL	15.144	6.561	23.727	***
ARGANZ – USERA	25.995	10.325	41.666	***
ARGANZ – VILLA	28.060	16.307	39.813	***

FUENCA – RETIRO	-49.798	-56.971	-42.625	***
FUENCA – CARABA	-31.753	-43.347	-20.160	***
FUENCA – SALAMA	-30.660	-36.400	-24.919	***
FUENCA – CHAMBE	-27.303	-33.691	-20.914	***
FUENCA – CENTRO	-24.572	-29.962	-19.182	***
FUENCA – CHAMAR	-3.284	-10.250	3.682	
FUENCA – BARAJA	-0.841	-7.274	5.593	
FUENCA – ARGANZ	-0.087	-7.681	7.506	
FUENCA – MONCLO	3.671	-2.769	10.112	
FUENCA – TETUAN	5.825	-0.697	12.346	
FUENCA - SAN BL	10.483	4.282	16.684	***
FUENCA – CIUDAD	12.785	5.809	19.761	***
FUENCA – HORTAL	15.056	6.693	23.420	***
FUENCA – USERA	25.908	10.357	41.459	***
FUENCA – VILLA	27.973	16.379	39.566	***
MONCLO – RETIRO	-53.469	-59.714	-47.224	***
MONCLO – CARABA	-35.424	-46.468	-24.381	***
MONCLO – SALAMA	-34.331	-38.858	-29.803	***
MONCLO – CHAMBE	-30.974	-36.299	-25.648	***
MONCLO – CENTRO	-28.243	-32.317	-24.169	***
MONCLO – CHAMAR	-6.955	-12.961	-0.949	***
MONCLO – BARAJA	-4.512	-9.890	0.867	
MONCLO – ARGANZ	-3.758	-10.482	2.965	
MONCLO – FUENCA	-3.671	-10.112	2.769	
MONCLO – TETUAN	2.153	-3.331	7.638	
MONCLO - SAN BL	6.812	1.713	11.910	***
MONCLO – CIUDAD	9.114	3.096	15.131	***
MONCLO – HORTAL	11.385	3.803	18.967	***
MONCLO – USERA	22.237	7.091	37.383	***
MONCLO – VILLA	24.301	13.258	35.345	***
TETUAN – RETIRO	-55.622	-61.951	-49.294	***
TETUAN – CARABA	-37.578	-48.668	-26.487	***
TETUAN – SALAMA	-36.484	-41.126	-31.842	***
TETUAN – CHAMBE	-33.127	-38.551	-27.704	***
TETUAN – CENTRO	-30.397	-34.597	-26.196	***
TETUAN – CHAMAR	-9.108	-15.201	-3.016	***
TETUAN – BARAJA	-6.665	-12.141	-1.190	***
TETUAN – ARGANZ	-5.912	-12.713	0.890	
TETUAN – FUENCA	-5.825	-12.346	0.697	
TETUAN – MONCLO	-2.153	-7.638	3.331	
TETUAN - SAN BL	4.658	-0.542	9.859	
TETUAN – CIUDAD	6.960	0.856	13.064	***
TETUAN – HORTAL	9.232	1.581	16.883	***
TETUAN – USERA	20.084	4.903	35.264	***
TETUAN – VILLA	22.148	11.057	33.239	***
SAN BL – RETIRO	-60.281	-66.279	-54.283	***
SAN BL – CARABA	-42.236	-53.141	-31.331	***
SAN BL – SALAMA	-41.143	-45.322	-36.963	***
SAN BL – CHAMBE	-37.786	-42.819	-32.752	***
SAN BL – CENTRO	-35.055	-38.738	-31.371	***
SAN BL – CHAMAR	-13.767	-19.515	-8.018	***
SAN BL – BARAJA	-11.323	-16.413	-6.234	***
SAN BL – ARGANZ	-10.570	-17.065	-4.075	***
SAN BL – FUENCA	-10.483	-16.684	-4.282	***
SAN BL – MONCLO	-6.812	-11.910	-1.713	***
SAN BL – TETUAN	-4.658	-9.859	0.542	
SAN BL – CIUDAD	2.302	-3.459	8.062	
SAN BL – HORTAL	4.574	-2.806	11.953	
SAN BL – USERA	15.425	0.380	30.471	***
SAN BL – VILLA	17.490	6.585	28.395	***
CIUDAD – RETIRO	-62.583	-69.379	-55.787	***

CIUDAD – CARABA	-44.538	-55.902	-33.174	***
CIUDAD – SALAMA	-43.444	-48.706	-38.183	***
CIUDAD – CHAMBE	-40.087	-46.049	-34.125	***
CIUDAD – CENTRO	-37.357	-42.233	-32.480	***
CIUDAD – CHAMAR	-16.069	-22.645	-9.492	***
CIUDAD – BARAJA	-13.625	-19.635	-7.616	***
CIUDAD – ARGANZ	-12.872	-20.110	-5.634	***
CIUDAD – FUENCA	-12.785	-19.761	-5.809	***
CIUDAD – MONCLO	-9.114	-15.131	-3.096	***
CIUDAD – TETUAN	-6.960	-13.064	-0.856	***
CIUDAD - SAN BL	-2.302	-8.062	3.459	
CIUDAD – HORTAL	2.272	-5.770	10.314	
CIUDAD – USERA	13.123	-2.258	28.504	
CIUDAD – VILLA	15.188	3.824	26.552	***
HORTAL – RETIRO	-64.854	-73.068	-56.641	***
HORTAL – CARABA	-46.810	-59.074	-34.545	***
HORTAL – SALAMA	-45.716	-52.713	-38.719	***
HORTAL – CHAMBE	-42.359	-49.897	-34.821	***
HORTAL – CENTRO	-39.628	-46.341	-32.916	***
HORTAL – CHAMAR	-18.340	-26.373	-10.307	***
HORTAL – BARAJA	-15.897	-23.473	-8.321	***
HORTAL – ARGANZ	-15.144	-23.727	-6.561	***
HORTAL – FUENCA	-15.056	-23.420	-6.693	***
HORTAL – MONCLO	-11.385	-18.967	-3.803	***
HORTAL – TETUAN	-9.232	-16.883	-1.581	***
HORTAL - SAN BL	-4.574	-11.953	2.806	
HORTAL – CIUDAD	-2.272	-10.314	5.770	
HORTAL – USERA	10.852	-5.206	26.909	
HORTAL – VILLA	12.916	0.652	25.180	***
USERA – RETIRO	-75.706	-91.177	-60.235	***
USERA – CARABA	-57.661	-75.614	-39.708	***
USERA – SALAMA	-56.568	-71.429	-41.706	***
USERA – CHAMBE	-53.211	-68.334	-38.087	***
USERA – CENTRO	-50.480	-65.210	-35.751	***
USERA – CHAMAR	-29.192	-44.568	-13.815	***
USERA – BARAJA	-26.749	-41.891	-11.606	***
USERA – ARGANZ	-25.995	-41.666	-10.325	***
USERA – FUENCA	-25.908	-41.459	-10.357	***
USERA – MONCLO	-22.237	-37.383	-7.091	***
USERA – TETUAN	-20.084	-35.264	-4.903	***
USERA - SAN BL	-15.425	-30.471	-0.380	***
USERA – CIUDAD	-13.123	-28.504	2.258	
USERA – HORTAL	-10.852	-26.909	5.206	
USERA – VILLA	2.065	-15.889	20.018	
VILLA – RETIRO	-77.771	-89.257	-66.285	***
VILLA – CARABA	-59.726	-74.384	-45.067	***
VILLA – SALAMA	-58.632	-69.282	-47.982	***
VILLA – CHAMBE	-55.275	-66.288	-44.262	***
VILLA – CENTRO	-52.545	-63.010	-42.079	***
VILLA – CHAMAR	-31.256	-42.614	-19.899	***
VILLA – BARAJA	-28.813	-39.852	-17.774	***
VILLA – ARGANZ	-28.060	-39.813	-16.307	***
VILLA – FUENCA	-27.973	-39.566	-16.379	***
VILLA – MONCLO	-24.301	-35.345	-13.258	***
VILLA – TETUAN	-22.148	-33.239	-11.057	***
VILLA - SAN BL	-17.490	-28.395	-6.585	***
VILLA – CIUDAD	-15.188	-26.552	-3.824	***
VILLA – HORTAL	-12.916	-25.180	-0.652	***
VILLA – USERA	-2.065	-20.018	15.889	

Tabla 14

7.2. Anexos Analíticos

7.2.1. Regresión

A continuación se presentan unos gráficos que recogen el comportamiento de la variable dependiente, precio, dependiendo, individualmente, de cada una de las variables.
Así, se podrá observar lo que en un principio parece lógico. Por ejemplo, los precios, en general, se abaratan cuando están sujetos a ofertas, o se encarecen a medida que la zona en que se encuentre cada hotel es "mejor" (por supuesto, todo esto está sujeto a excepciones aisladas). En los casos de las variables *CPPROM*, *FECHAPROM*, *OFERTAPROM* y *ESTRELLASPROM*, originalmente de cualitativas, en vez de graficar el precio promedio para cada nivel con que hemos graficado, se mostrará la variación del precio con respecto a sus distintas categorías.

A. Distribución del precio atendiendo a las distintas variables del mejor modelo de regresión.

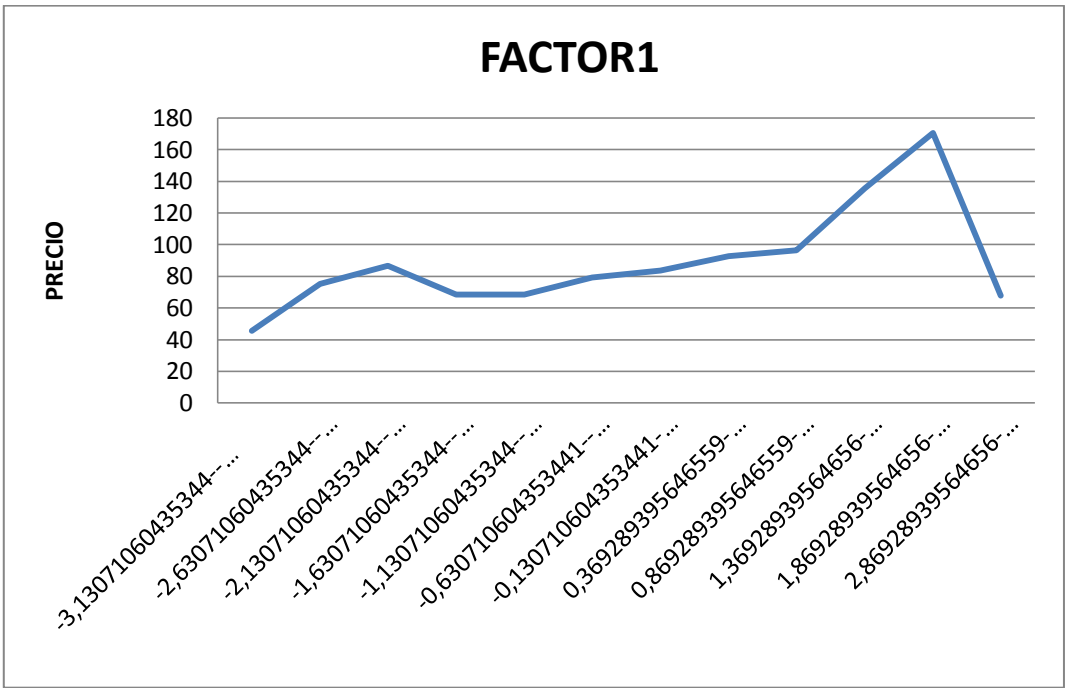

Figura 27. Distribución del precio con respecto al comportamiento del factor 1.

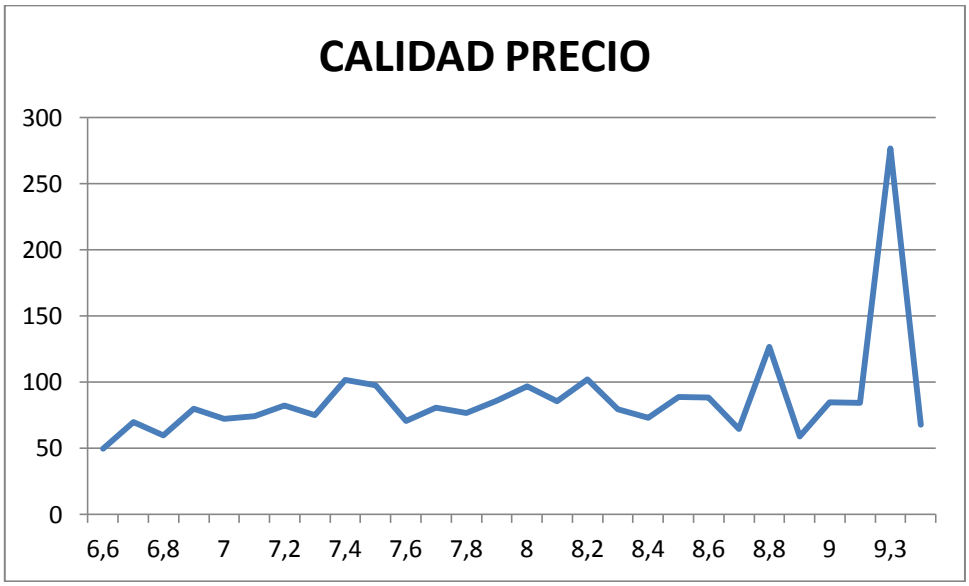

Figura 28. Distribución del precio con respecto a la valoración de la relación calidad-precio

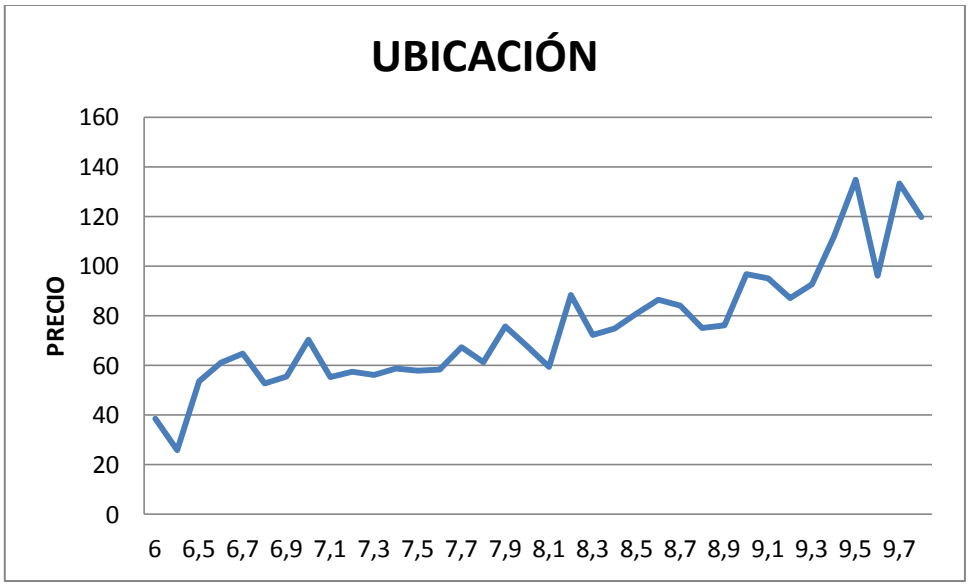

Figura 29. Distribución del precio con respecto a la valoración de la ubicación.

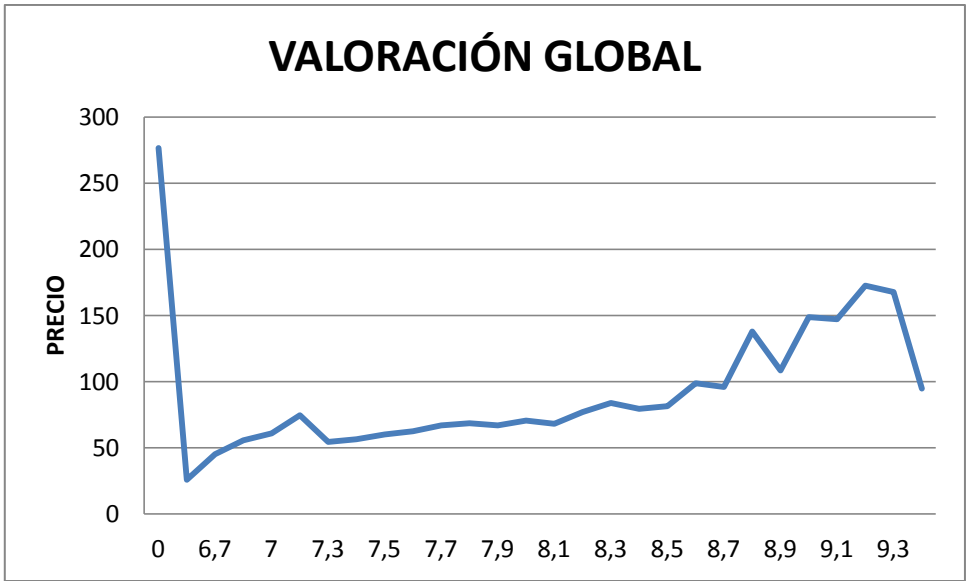

Figura 30. Distribución del precio con respecto a la valoración global.

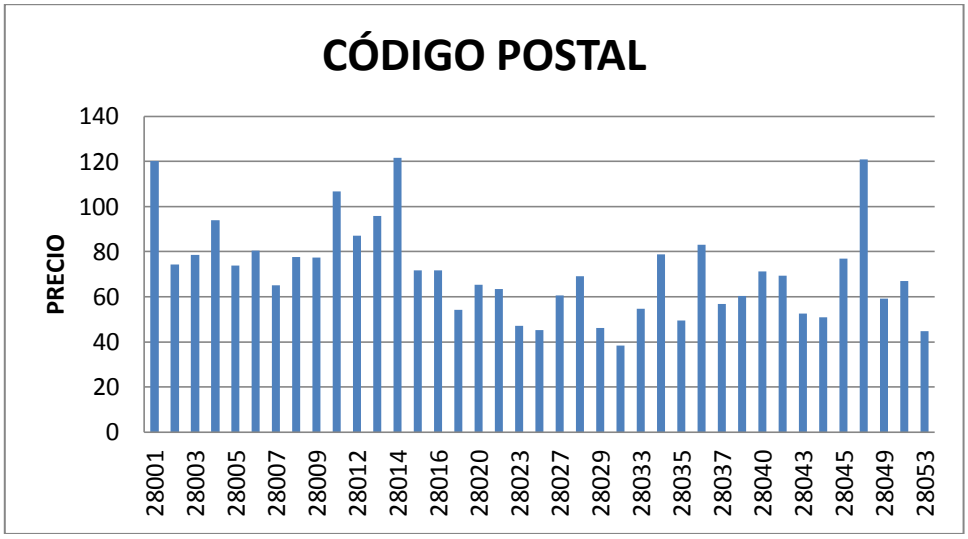

Figura 31. Distribución del precio con respecto al código postal.

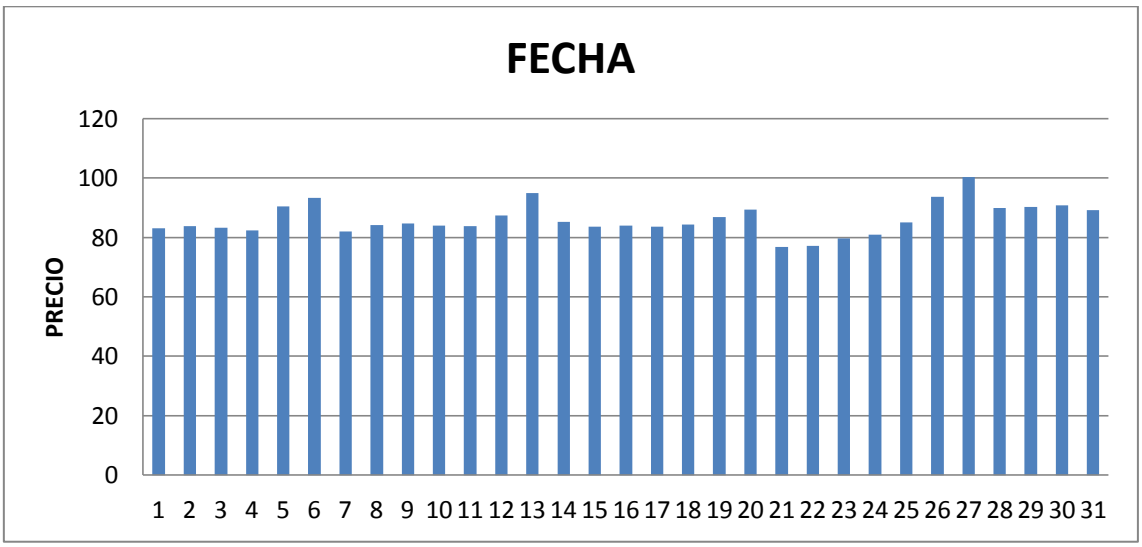

Figura 32. Distribución del precio con respecto a la fecha.

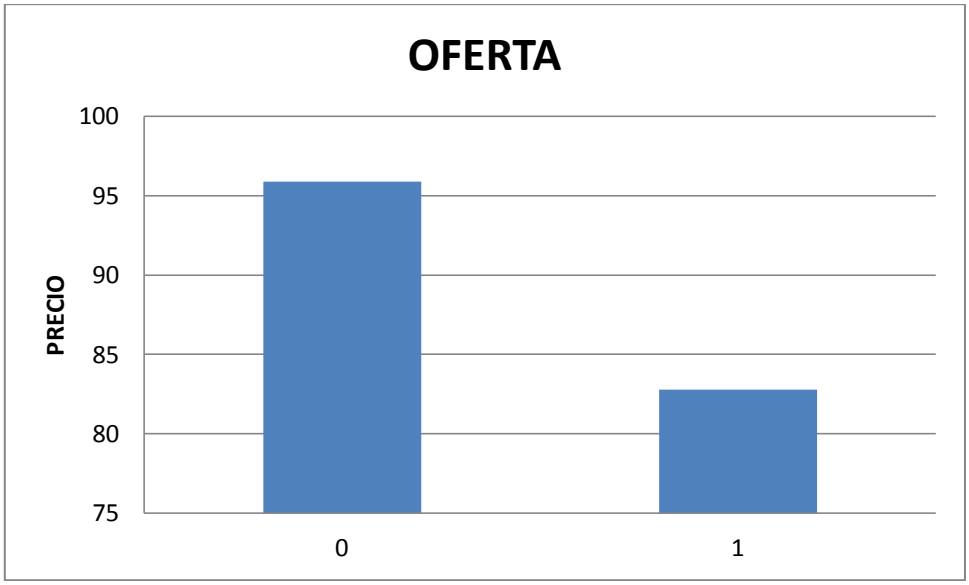

Figura 33. Distribución del precio atendiendo si es una oferta o no

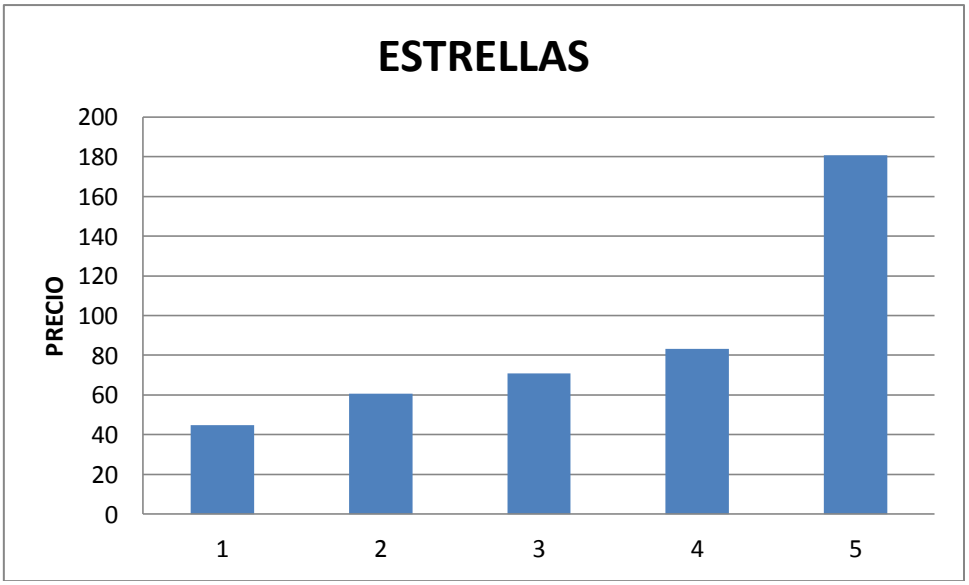

Figura 34. Distribución del precio atendiendo a las estrellas.

B. Transformación de las variables del mejor modelo de regresión.

FACTOR 1

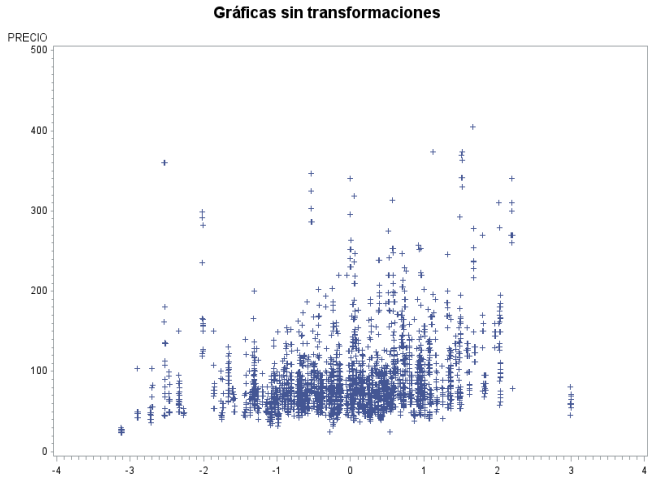

Figura 35. Sin transformación

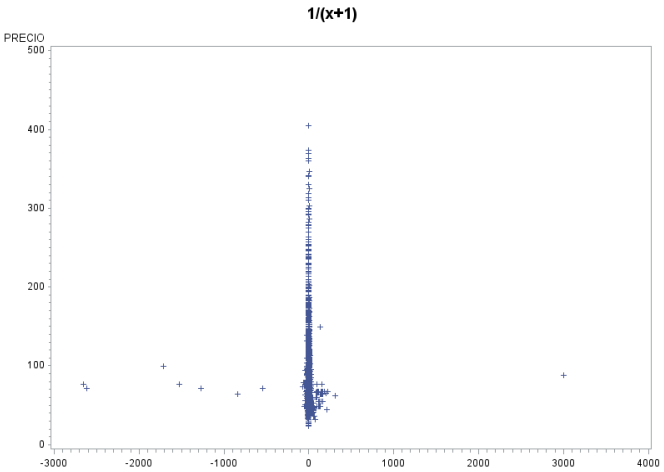

Figura 36. 1/(x+1)

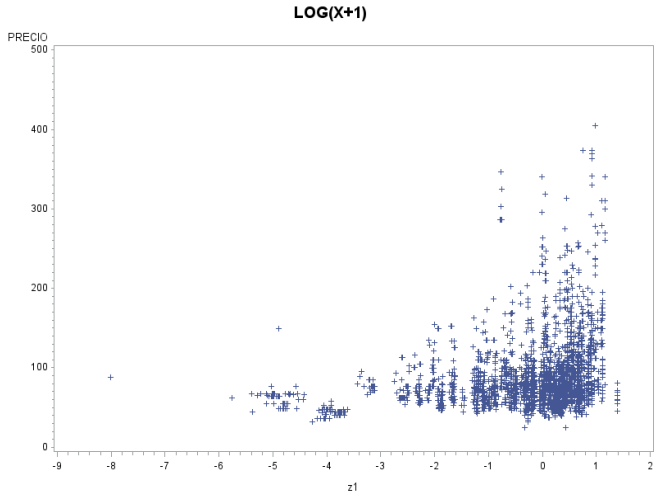

Figura 37. Log(x+1)

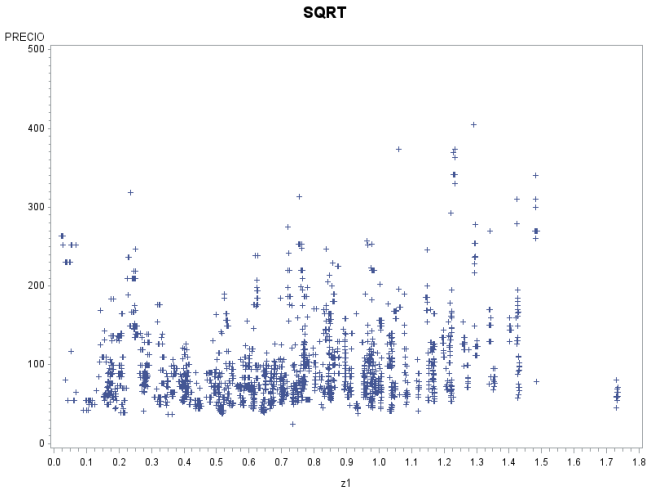

Figura 38. SQRT

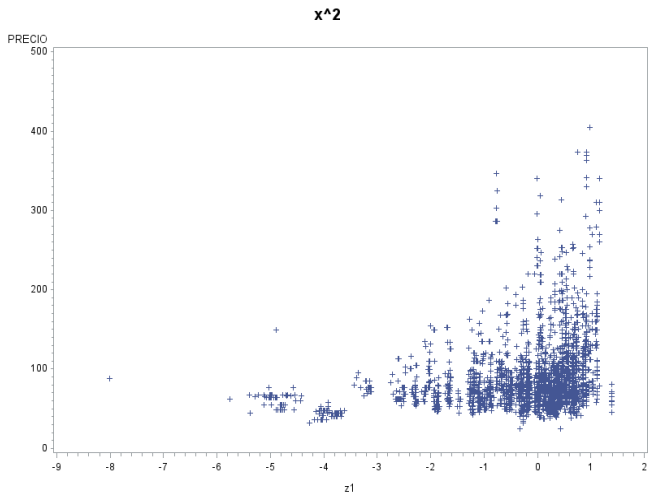

Figura 39. x^2

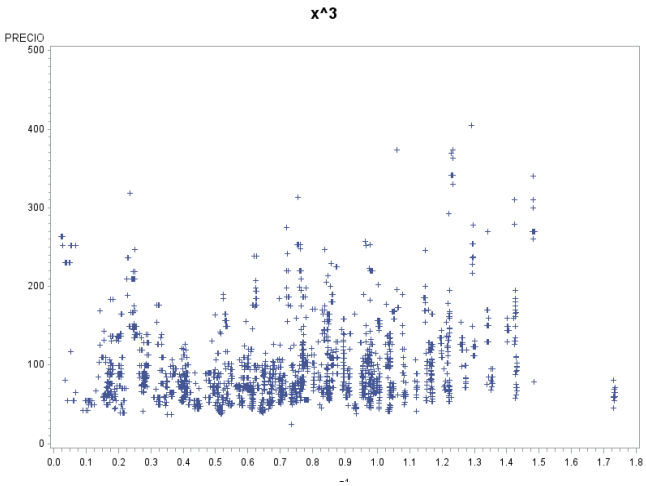

Figura 40. x^3

CALIDAD PRECIO

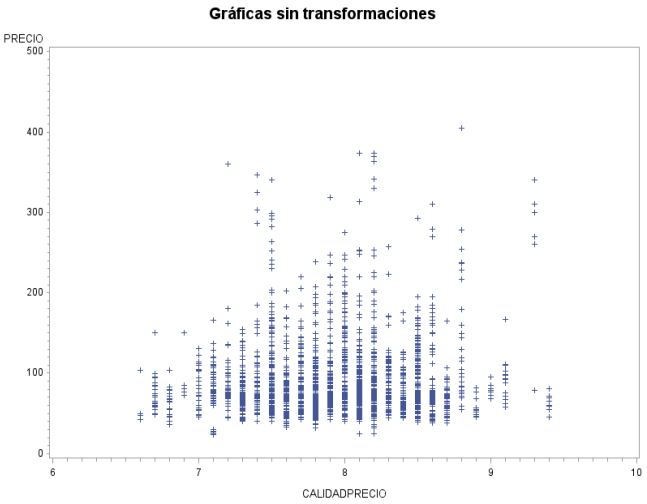

Figura 41. Sin transformación

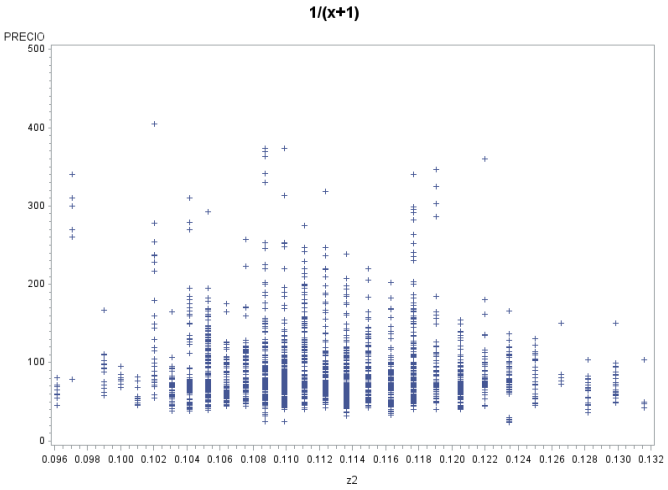

Figura 42. 1/((x+1)

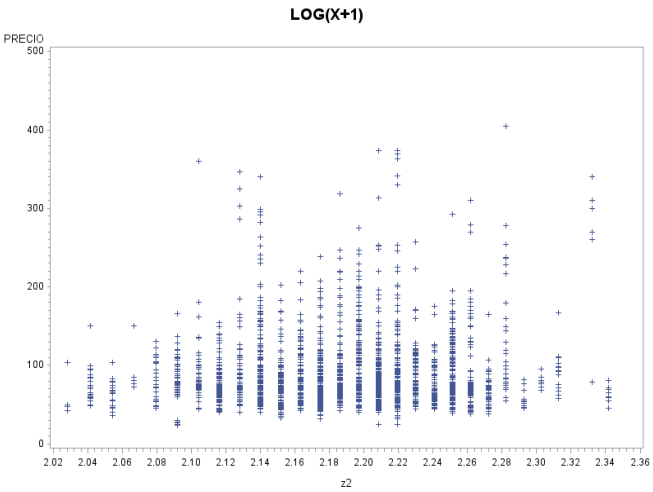

Figura 43. Log(x+1)

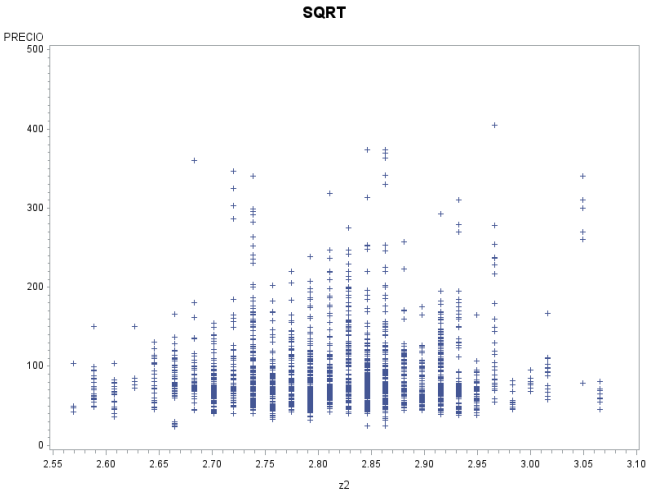

Figura 44. SQRT

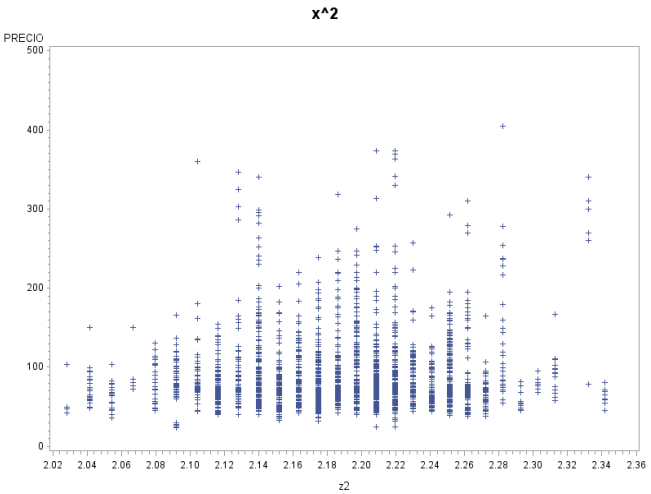

Figura 45. x^2

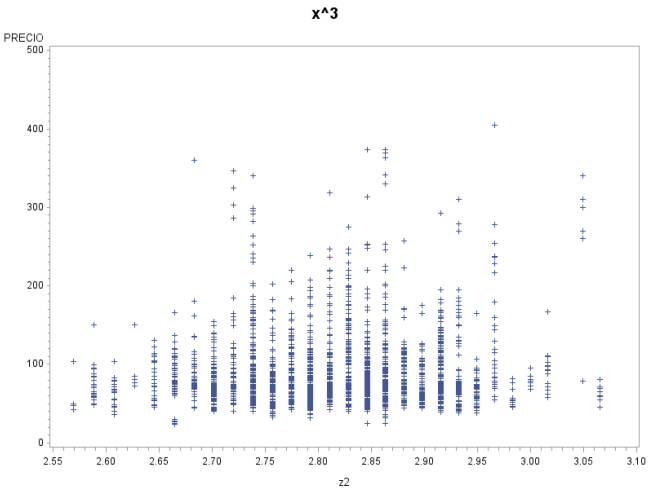

Figura 46. x^3

UBICACIÓN

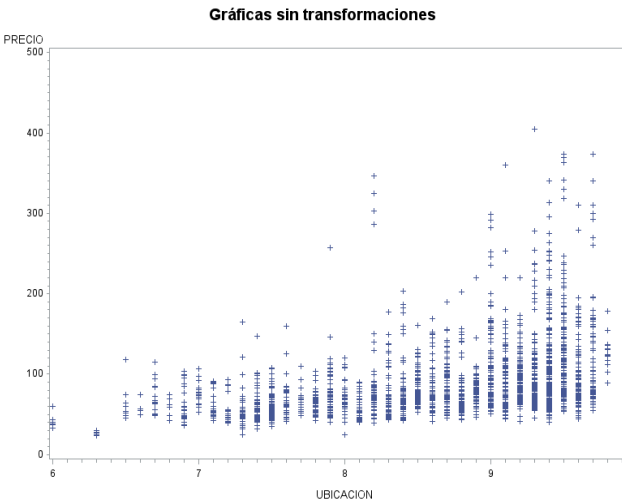

Figura 47. Sin transformación

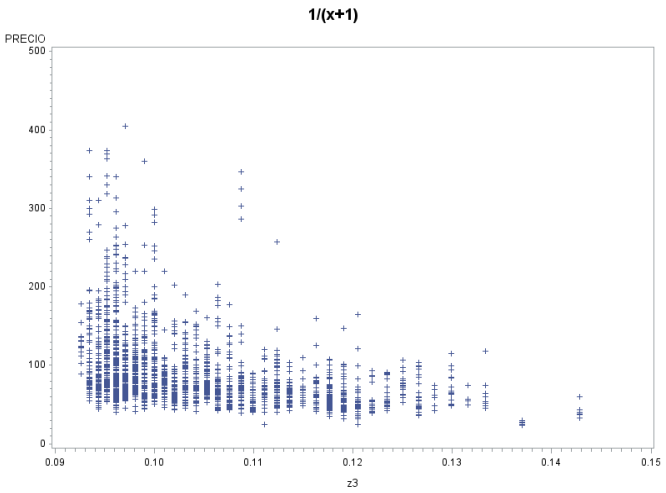

Figura 48. 1/(x+1)

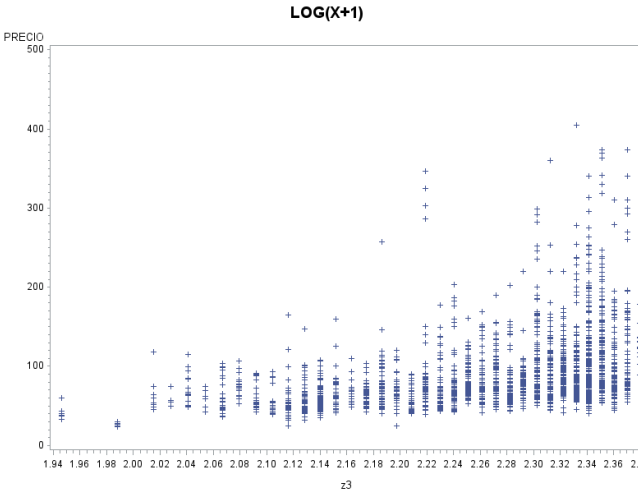

Figura 49. Log(x+1)

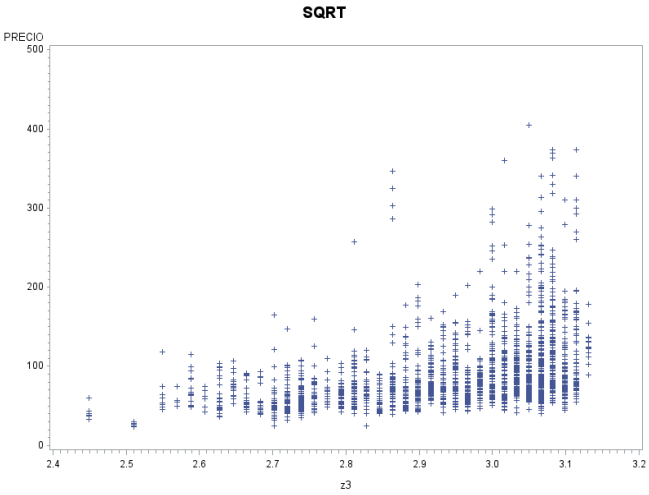

Figura 50. SQRT

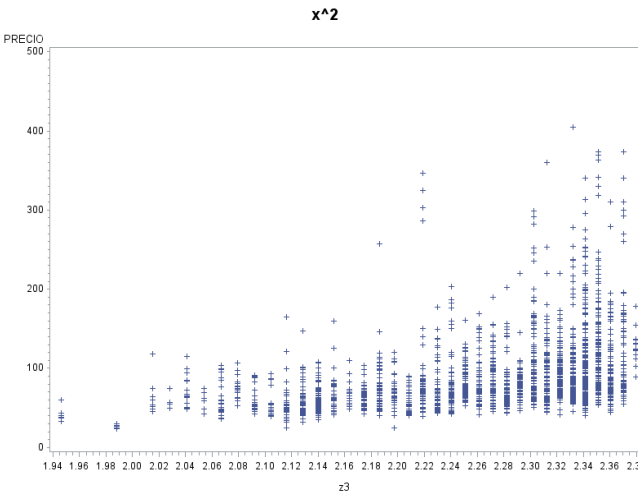

Figura 51. x^2

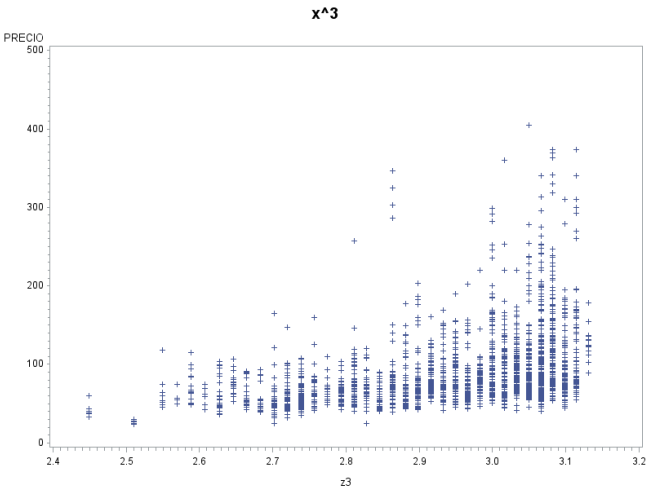

Figura 52. x^3

VALORACIÓN GLOBAL

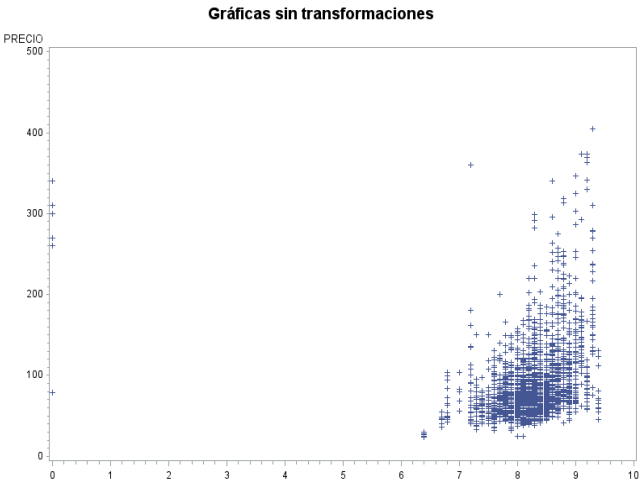

Figura 53. Sin transformación

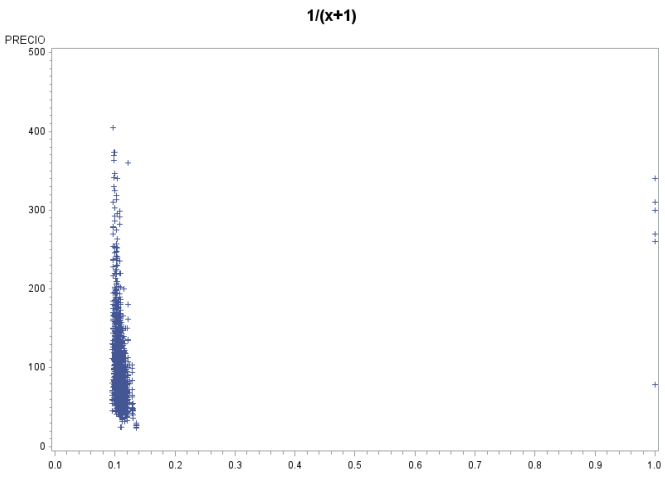

Figura 54. 1/(x+1)

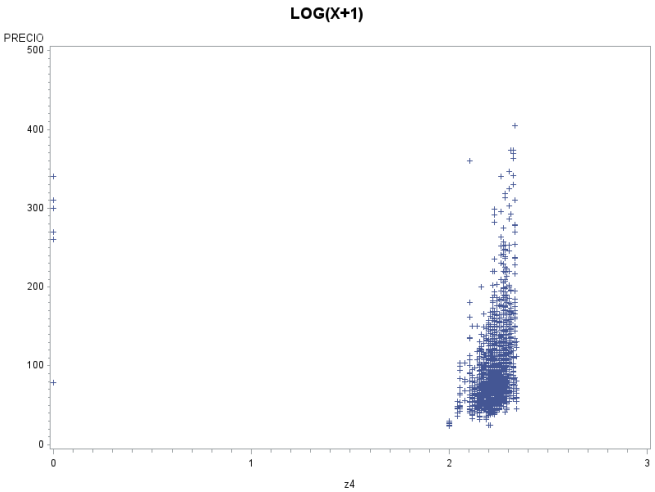

Figura 55. Log(x+1)

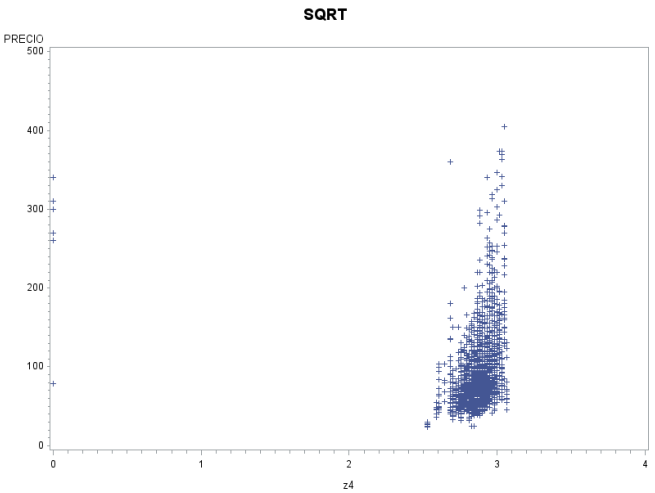

Figura 56. SQRT

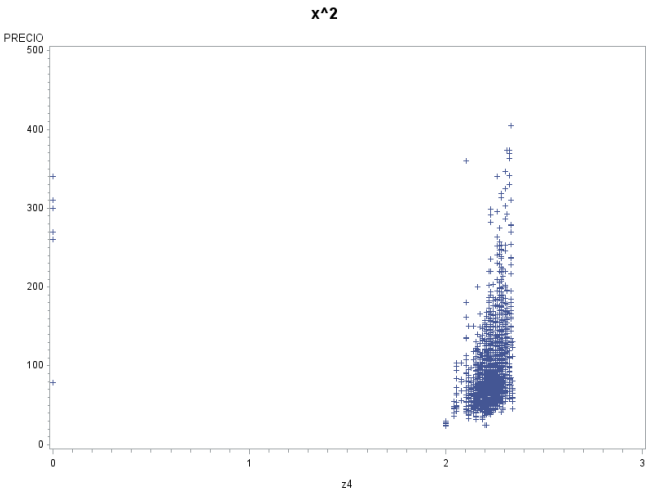

Figura 57. x^2

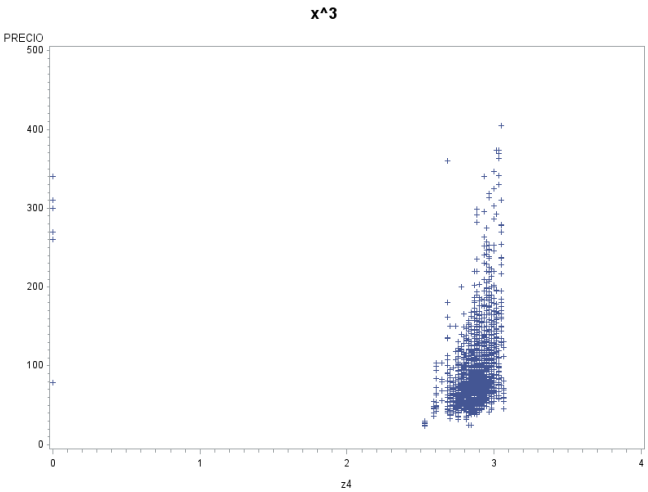

Figura 58. x^3

CPPROM (Promedio del precio para los distintos códigos postales)

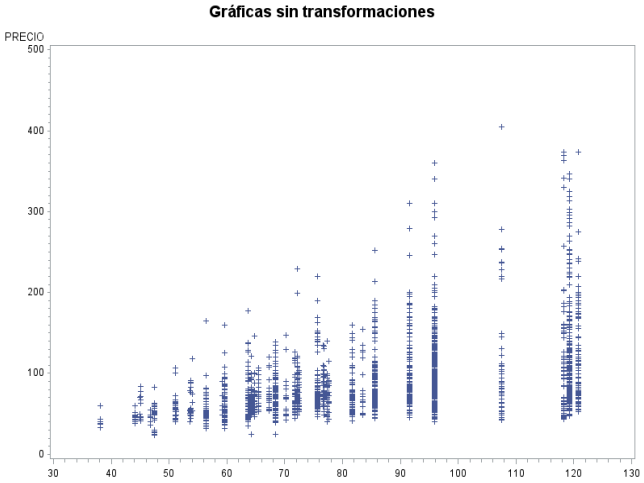

Figura 59. Sin transformación.

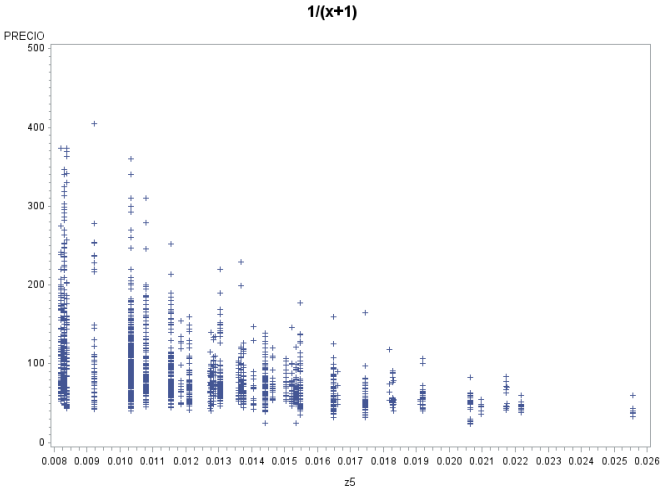

Figura 60 . 1/(x+1)

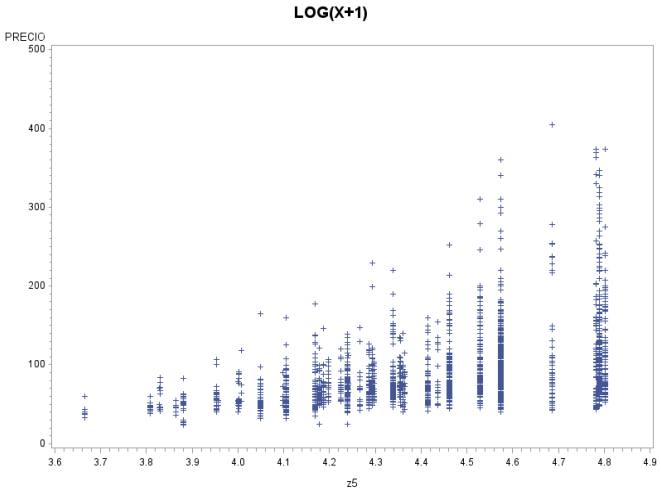

Figura 61. Log(x+1)

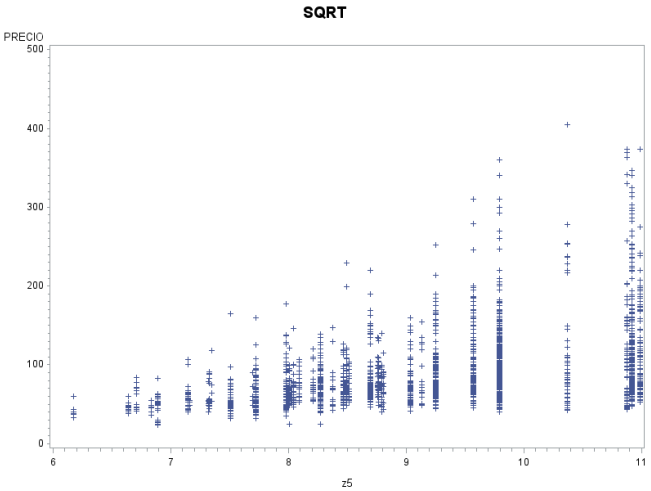

Figura 62 SQRT

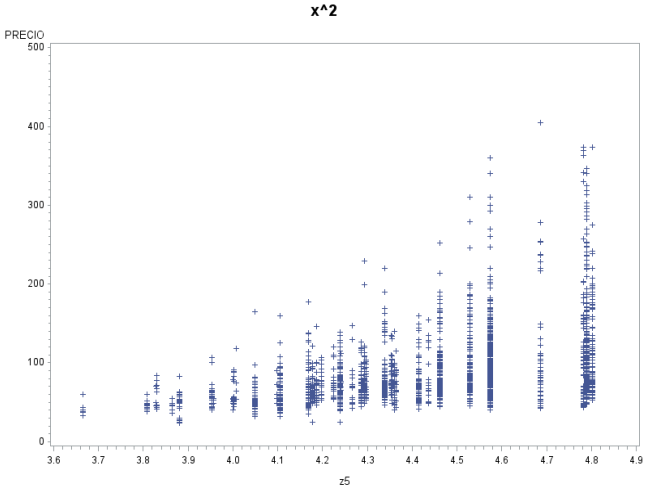

Figura 63. x^2

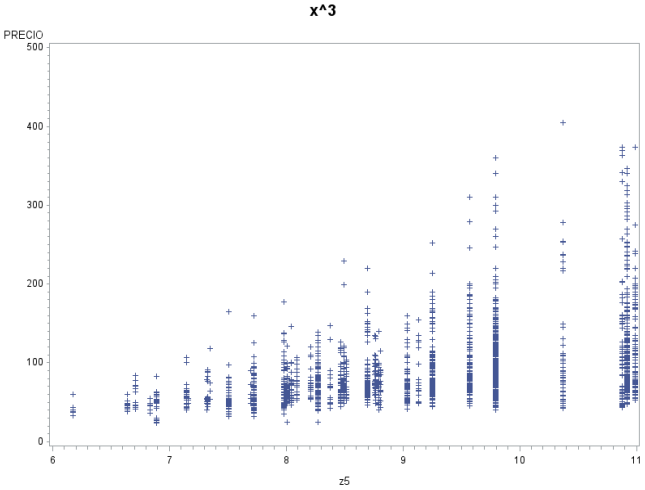

Figura 64. x^3

FECHAPROM (Promedio del precio para las distintas fechas)

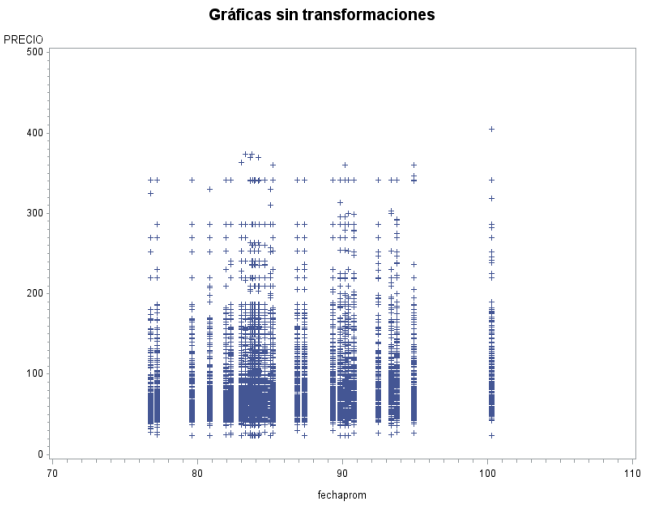

Figura 65. Sin transformación

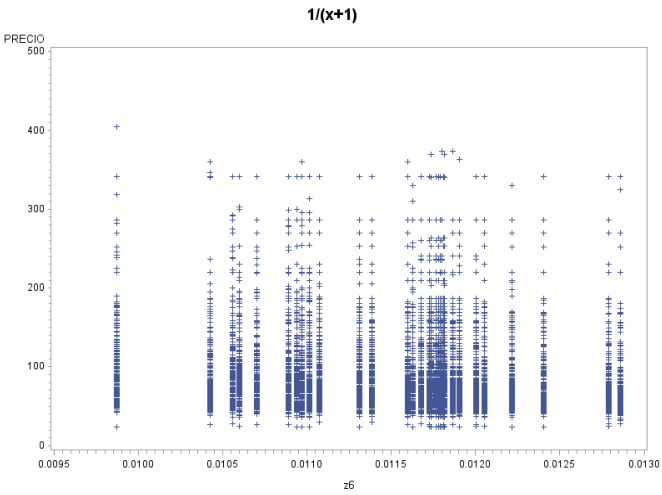

Figura 66. 1/(x+1)

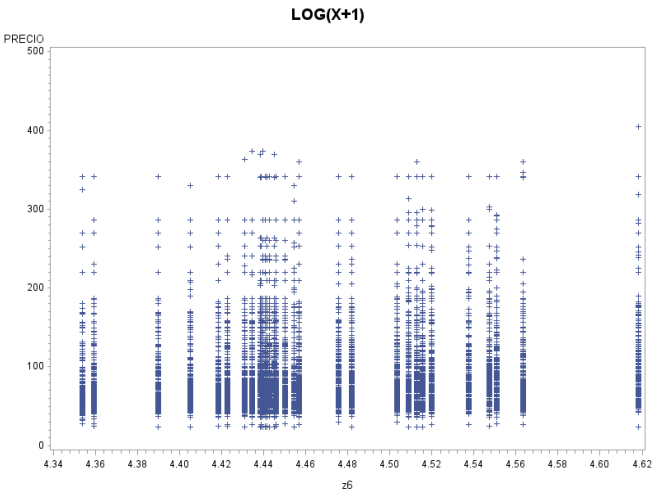

Figura 67. Log(x+1)

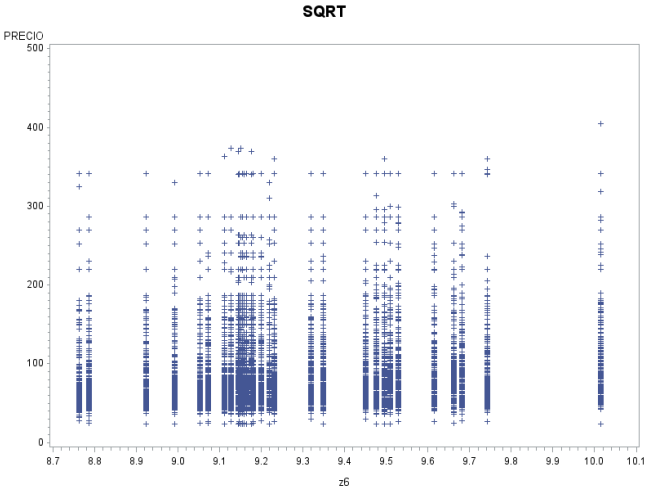

Figura 68. SQRT

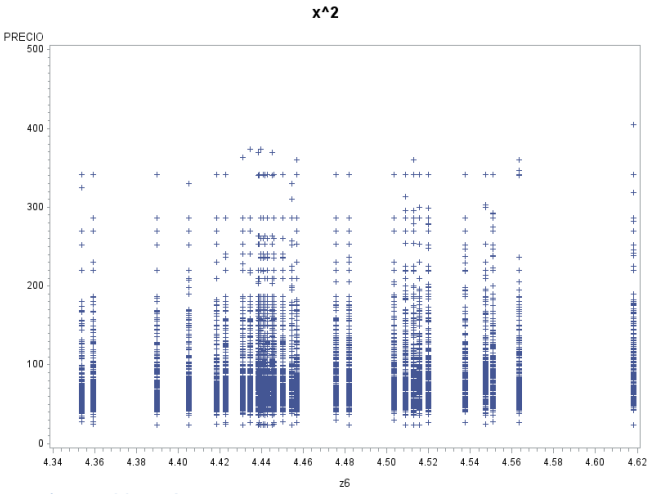

Figura 69. x^2

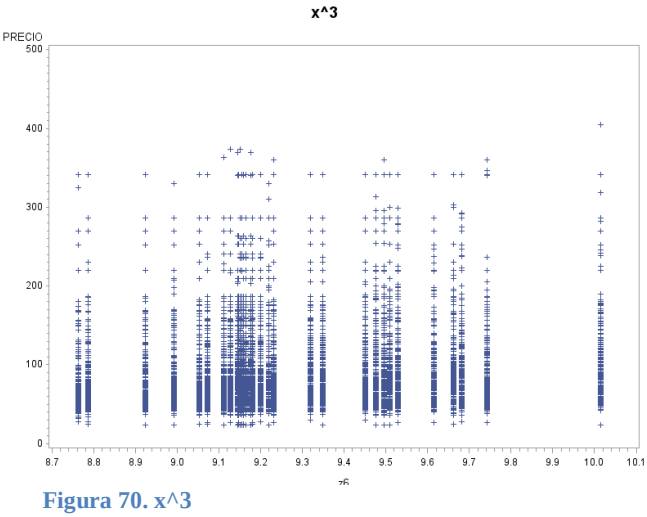

Figura 70. x^3

OFERTAPROM (Promedio del precio con respecto a la presencia de ofertas)

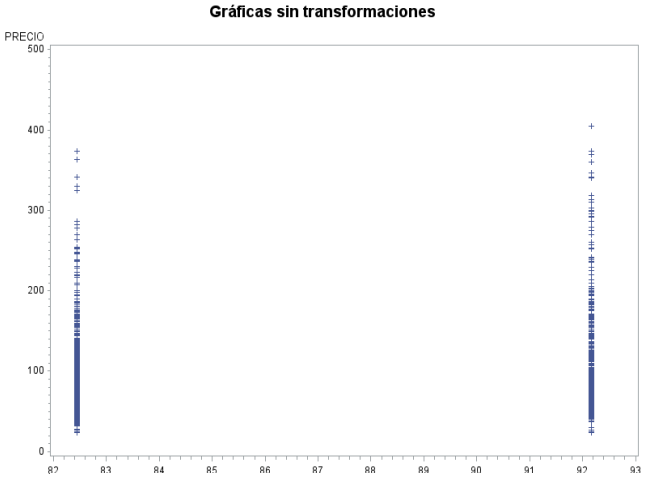

Figura 71. Sin transformación

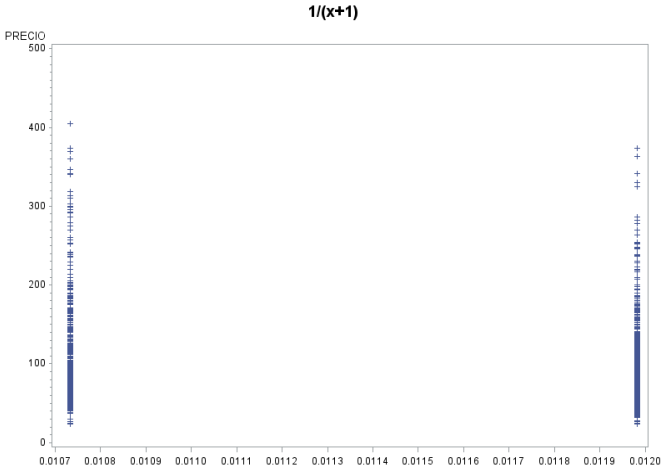

Figura 72. 1/(x+1)

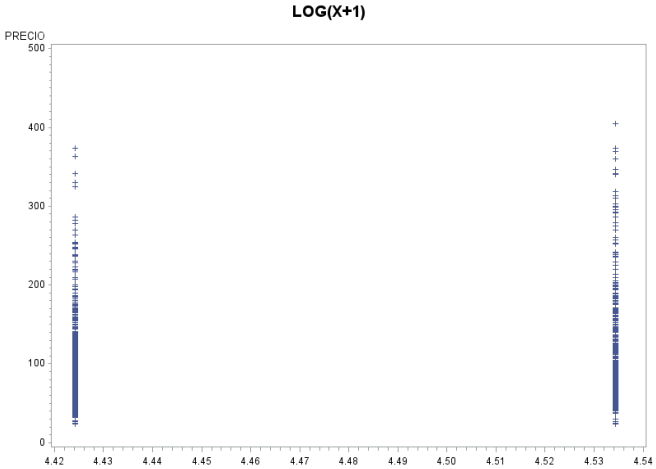

Figura 73. Log(x+1)

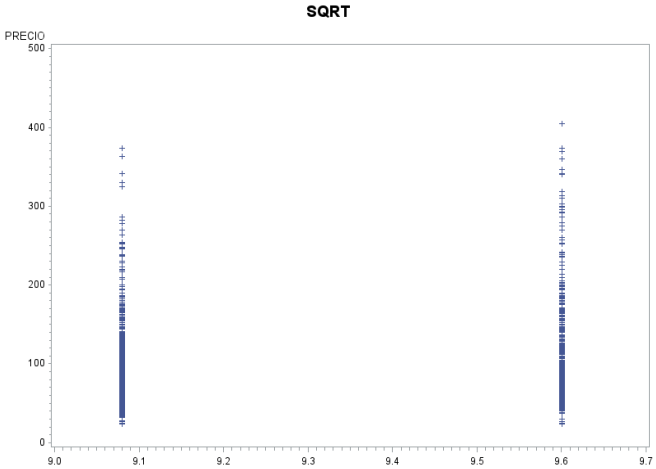

Figura 74. SQRT

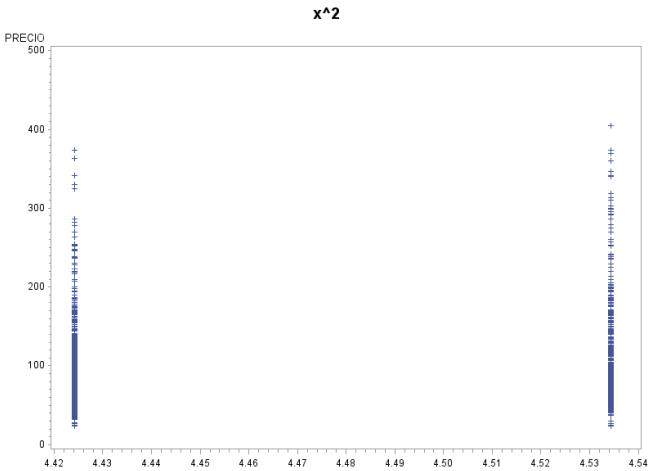

Figura 75. x^2

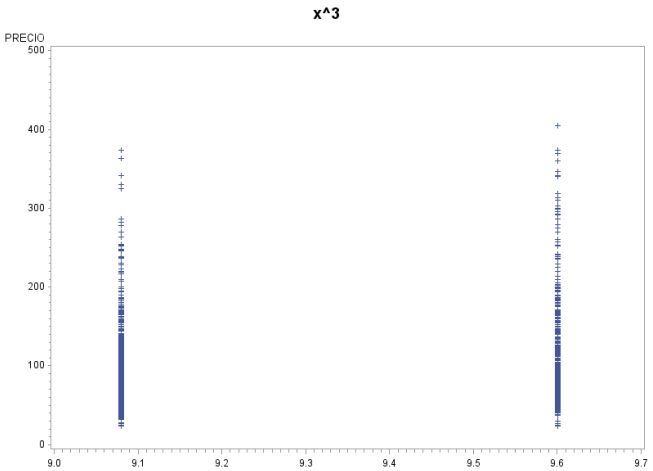

Figura 76. x^3

ESTRELLASPROM (Promedio del precio con respecto a las estrellas)

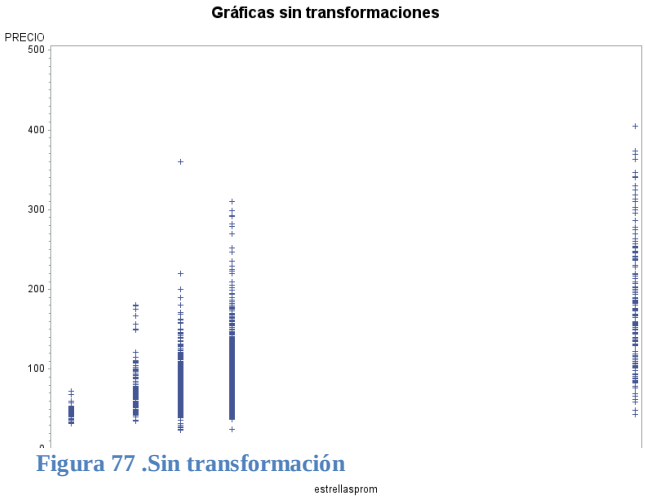

Figura 77 .Sin transformación

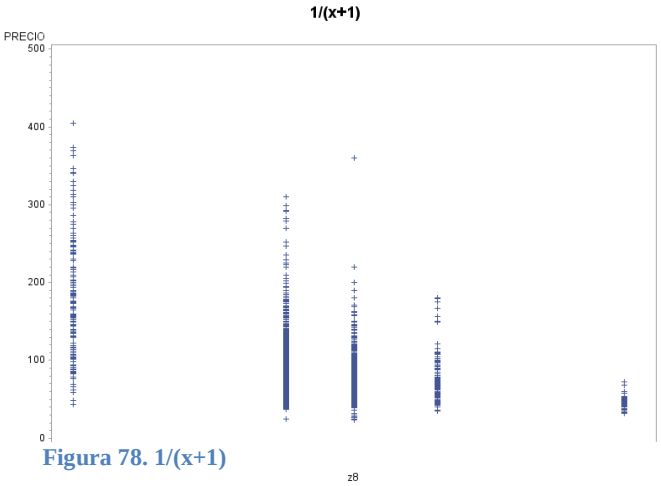

Figura 78. 1/(x+1)

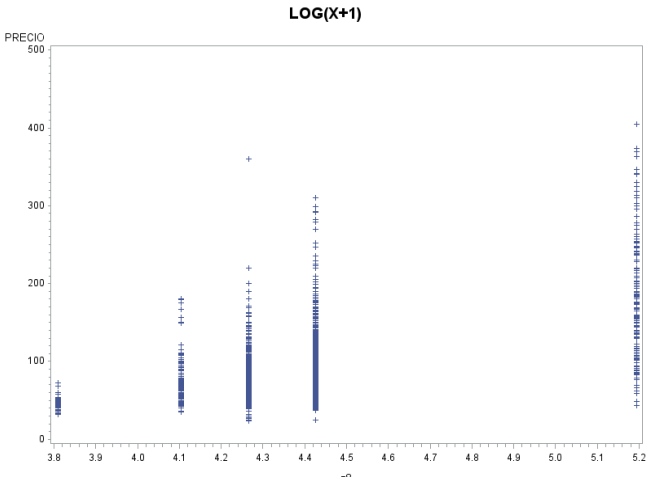

Figura 79. Log(x+1)

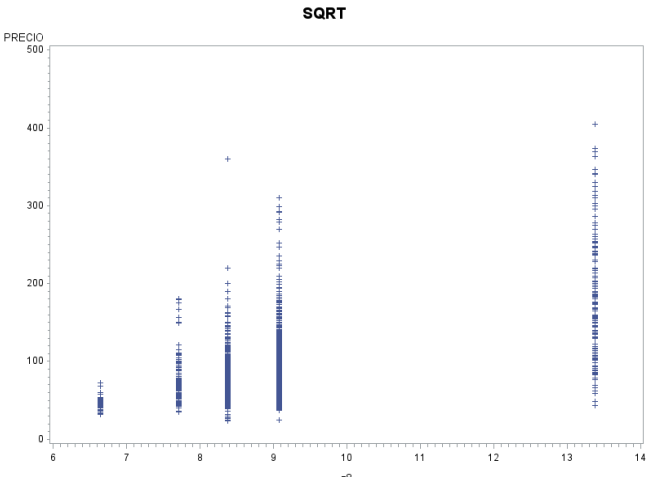

Figura 80. SQRT

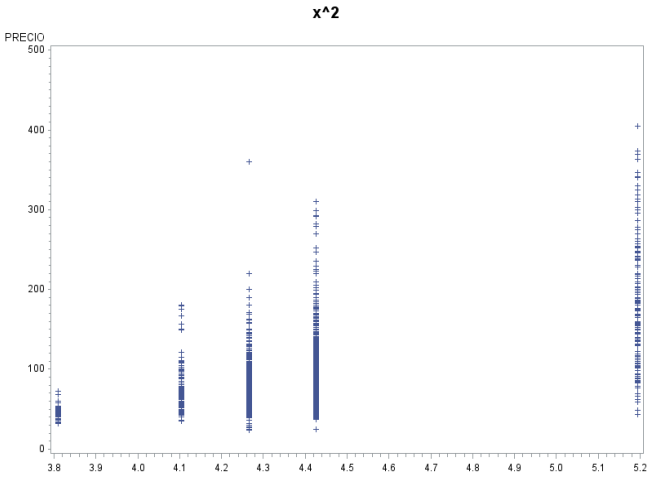

Figura 81.x^2

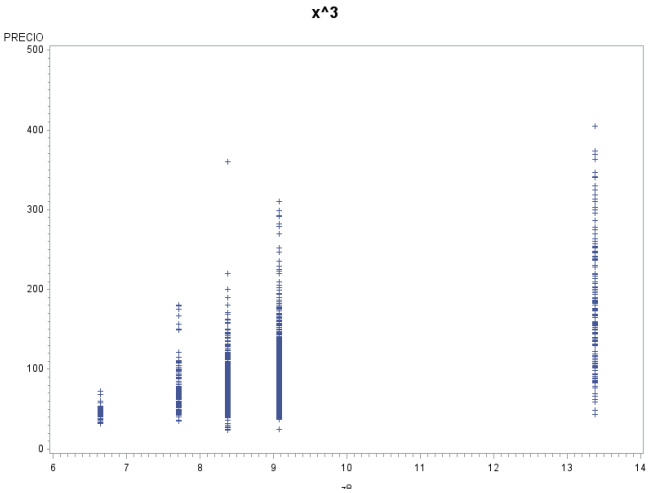

Figura 82. x^3

7.3. Código SAS Base

```
/*----------------------ANALISIS FACTORIAL--------------------*/
procfactordata=discoc.base          out=discoc.factorial1valoracionesok
outstat=discoc.factorial2valoracionesok priors=smc nfactors=1;
var   VALGLOBAL   LIMPIEZA   CONFORT   INSTALACIONESSERVICIOS   PERSONAL
CALIDADPRECIO WIFI
;
run;

/*rotate= VARIMAX, QUARTIMAX, PARSIMAX*/
procfactordata=discoc.base              out=discoc.factorial1todasparsi
outstat=discoc.factorial2todasparsi                          priors=smc
nfactors=19ROTATE=parsimax;
var  OFERTA  FECHA  ESTRELLAS  FECHAINSC  COMENTARIOS  VALGLOBAL  LIMPIEZA
CONFORT UBICACION INSTALACIONESSERVICIOS PERSONAL
CALIDADPRECIO WIFI USABLE ESTACIONAL FINDE CPPROM zonaprom diaprom
;
run;

/*-------------CLUSTER JERARQUICO-------------------------*/

PROCACECLUSDATA=discoc.factorial1todasvarimax OUT=discoc.F P=.03 ;
VAR factor1 factor2 factor3;
RUN;
/*METHOD=AVERAGE, WARD*/
proccluster data=discoc.factorial1todasvarimax            method=average
pseudoouttree=uno;
var factor1 factor2 factor3;run;
proctreedata  =  uno  pos=20  OUT=discoc.NEW  N=8  GRAPHICS  HAXIS=AXIS1
HORIZONTAL; run;

/*-------------------CLUSTER NO JERARQUICO------------------*/

procfastclusdata=discoc.factorial1todasvarimax  maxclusters=8maxiter=10
out=discoc.cluster;
var factor1 factor2 factor3;
run;


/*---------------------------REGRESION----------------*/
/*------------1. OBTENCION DE MODELOS TENTATIVOS INICIALES---------*/

odsoutput SelectedEffects=efectos;
procglmselectdata      =      discoc.train3      valdata=discoc.valida3
testdata=discoc.test3;
class;
model  precio=CALIDADPRECIO  CONFORT  INSTALACIONESSERVICIOS  LIMPIEZA
PERSONAL UBICACION VALGLOBAL WIFI
CPPROM estrellasprom zonaprom diaprom ofertaprom fechaprom H_PROM
/selection=backward /*slentry=0.5*/ slstay=0.1;/*el de salida siempre
>=entrada EN STEPWISE!!! en backward solo criterio de salida y en
forward de enrtada*/
scoredata=discoc.train3 out=salpredi RESIDUAL=RESIDUO;
scoredata=discoc.valida3 out=salpredival RESIDUAL=RESIDUO;
scoredata=discoc.test3 out=salpreditest RESIDUAL=RESIDUO;
run;
data;set efectos; put effects @@;run;


procfactordata=discoc.base3                          out=discoc.factorial
outstat=discoc.factorial2todo3 priors=smc nfactors=11ROTATE=VARIMAX;
```

```
var CALIDADPRECIO CONFORT INSTALACIONESSERVICIOS LIMPIEZA PERSONAL
UBICACION VALGLOBAL WIFI FINDE CPPROM ZONAPROM DIAPROM FECHAPROM
OFERTAPROM ESTRELLASPROM
;run;
/*--HACEMOS VALIDACIÓN CRUZADA, USANDO ARCHIVOS TRAINING Y VALIDATION-
--*/

%trainvalida(archivo=base,vardepen=precio,listclass=H            OFERTA
FECHA,listindepen=H OFERTA FECHA,
porcen=0.68,seminicio=12445,semifinal=12500);
data final101;set final;modelo=101;
data discoc.final101;set final101;run;
%macro
trainvalida(archivo=,vardepen=,listclass=,listindepen=,porcen=,seminic
io=,semifinal=);
data final;run;
%do  semilla=&seminicio  %to&semifinal;/*<<<<<******AQUI  SE  PUEDEN
CAMBIAR LAS SEMILLAS */
data doss;set &archivo;u=ranuni(&semilla);
proc sort data=doss;by u;run;

data doss ;
set doss nobs=nume;
if _n_<=&porcen*nume then vardep=&vardepen;else vardep=.;
run;

%if&listclass ne %then%do;
/*************************************************************/
proc glm data=doss noprint;/*<<<<<******SE PUEDE QUITAR EL NOPRINT */
class &listclass;
model vardep=&listindepen;
output out=sal p=predi;run;
/*************************************************************/
data   sal;set   sal;resi2=(&vardepen-predi)**2;if   vardep=.   then
output;run;
%end;
%else%do;
/*************************************************************/
proc glm data=doss noprint;/*<<<<<******SE PUEDE QUITAR EL NOPRINT */
model vardep=&listindepen;
output out=sal p=predi;run;
/*************************************************************/
data   sal;set   sal;resi2=(&vardepen-predi)**2;if   vardep=.   then
output;run;
%end;

proc means data=sal noprint;var resi2;
output out=mediaresi mean=ASE;
run;
data mediaresi;set mediaresi;semilla=&semilla;media=ASE;run;

data final (keep=media semilla);set final mediaresi;if media=. then
delete;run;
%end;
proc print data=final;run;
%mend;



/*----------PASO 2 ESTUDIO DE LAS VARIABLES CATEGORICAS----------*/
```

```
/*como ya sabemos de los primeros pasoss en la investigación, hay
variables nominales que tienen alguna de sus categorias
muy mal representadas
con la macro agrupacategorias, vamos a intentar agrupar estas
variables
*/
optionnonotes;
%AgruparCategorias(
archivo=test,vardep=precio,vardeptipo=I,
listclass=H OFERTA FECHA ESTRELLAS FECHAINSC ZONA CODPOSTAL
COMENTARIOS USABLE ESTACIONAL DIA FINDE HOTEL, criterio=,
directorio=C:\Users\Inma Gutiérrez\Documents\MÁSTER\TFM);
%macro AgruparCategorias(
archivo=,         /* Archivo de datos que contiene a las variables
Nominales */
vardep=,          /* Variable Dependiente (Intervalo o Nominal ) */
vardeptipo=,      /* Tipo de la variable dependiente: I=Intervalo o
N=Nominal */
listclass=,       /* Lista separada por espacios de las variables a
agrupar */
criterio=,        /* Criterio usado para la división de las ramas en el
proc arboretum */
directorio=c:/* directorio de trabajo para archivos de apoyo */);
%if&criterio eq %then%do;
%if&vardeptipo=I %then%let criterio=PROBF;
%if&vardeptipo=N %then%let criterio=PROBCHISQ;
%end;
      /* Solo con la información relevante */
      data archivosa;
            set &archivo (KEEP = &vardep &listclass);
      run;
      data _null_;
            file
            "&directorio\tempAgrupacionClasesVariableNominal.txt";
            put ' ';
      run;
      data _null_;
            length clase $ 10000 ;
      /* Cuento el número de variables */
      clase="&listclass";
            ncate= 1;
            do while (scanq(clase, ncate) ^= '');
                  ncate+1;
            end;
            ncate+(-1);put;
            put // ncate= /;
            call symput('ncate',left(ncate));
      run;
      /* Bucle arboretum */
      %do i=1%to&ncate;
            %let vari=%qscan(&listclass,&i);
            %if%upcase(&vardeptipo)=I %then%do;
                  proc arboretum data=archivosa criterion=&criterio;
                  /* CRITERIO PROBF HACE CONTRASTES TIPO PARES */
                        input &vari / level=nominal;
                        target &vardep / level=interval;
                        save model=tree1;
                  run;
            %end;
            %else%do;
                  proc arboretum data=archivosa criterion=&criterio;
```

```
            input &vari / level=nominal;
            target &vardep / level=nominal;
            save model=tree1;
      run;
%end;
proc arboretum inmodel=tree1;
      score data=archivosa out=archivosa2 ;
      subtree best;
run;
data archivosa;
      set archivosa2;
run;
/* comprobar si no se hacen agrupaciones */
proc freq data=archivosa noprint;
      tables &vari /out=sal1;
proc freq data=archivosa noprint;
      tables _leaf_ /out=sal2;
data _null_;
      if _n_=1 then set sal1 nobs=nume1;
      if _n_=1 then set sal2 nobs=nume2;
      if _n_=1 then do;
            if nume1=nume2 then noagrupa=1;
            else noagrupa=0;
            call symput ('noagrupa',left(noagrupa));
      end;
      if noagrupa=1 then do;
            put 'NOAGRUPA  '"&vari";
            file
            "&directorio\tempAgrupacionClasesVariableNomin
            al.txt" mod;
            put "&vari";
      end;
      stop;
run;
/* comprobar si se unen todas las categorías */
proc freq data=archivosa noprint;
      tables _leaf_ /out=sal1;
run;
data _null_;
      set sal1 nobs=nume;
      call symput ('seunentodas',left(nume));
      if nume=1 then do;
            put 'SE UNEN TODAS   '"&vari";
            file
            "&directorio\tempAgrupacionClasesVariableNomin
            al.txt" mod;
            put "&vari";
      end;run;
%if&noagrupa eq 0 and &seunentodas ne 1%then%do;
      data            _null_;koko2=cats("&vari",'_G');call
      symput('koko',left(koko2));run;
      data archivosa (drop=_node_ );
            set archivosa;
            rename _leaf_=&koko;
      run;
      data _null_;
            file
            "&directorio\tempAgrupacionClasesVariableNomin
            al.txt" mod;
            h="&koko";h=left(h);
            put h;
```

```
                  run;
            %end;
            %else%do;
                  data archivosa(drop=_leaf_ _node_);
                        set archivosa;
                  run;
            %end;
      %end;
      data archivofinal (drop=P_&vardep R_&vardep);
            merge &archivo archivosa;
      run;
      data _null_;
            length c $ 300;
            if _n_=1 then put ' '//'LISTA DE GRUPOS CREADOS Y NO
            CREADOS'//'****************************************' ;
                  infile
                  "&directorio\tempAgrupacionClasesVariableNominal.txt
                  " ;
                  input c $;
                  put c @@;
      run;
      data                                                    _null_;put
      //'****************************************';run;

/* COMPROBAR GRUPOS CREADOS */
%do i=1%to&ncate;
      %let vari=%qscan(&listclass,&i);
      data    _null_;retain    control    0;length    c    $    300;infile
      "&directorio\tempAgrupacionClasesVariableNominal.txt" ;input c$;
            c3=cats("&vari",'_G');
            if c=c3 then control=1;
            call symput('control',left(control));
            call symput('grupo',left(c3));run;
      %if&control=1%then%do;
         proc   freq   data=archivofinal   noprint;tables   &vari*&grupo
         /out=sal;run;
      proc sort data=sal;by &grupo;
      proc print data=sal;run;
      %end;%end;  %mend;
%renombrar(archivo=base,listaclass=ZONA DIA HOTEL,
listaconti=CALIDADPRECIO   CONFORT   INSTALACIONESSERVICIOS   LIMPIEZA
PERSONAL UBICACION VALGLOBAL WIFI
H  OFERTA  FECHA  ESTRELLAS  FECHAINSC  CODPOSTAL  COMENTARIOS  USABLE
ESTACIONAL FINDE PRECIO,
prefijoclass=X,prefijoconti=Z);
%macro
renombrar(archivo=,listaclass=,listaconti=,prefijoclass=,prefijoconti)
;
%if&listaconti ne %then%do;
data _null_;
clase="&listaconti";
   nconti= 1;
   do while (scanq(clase, nconti) ^= '');
      nconti+1;
   end;
   nconti+(-1);
   call symput('nconti',left(nconti));run;
%end;
%if&listaclass ne %then%do;
data _null_;
clase="&listaclass";
```

```
   ncate= 1;
   do while (scanq(clase, ncate) ^= '');
      ncate+1;
   end;
   ncate+(-1);
   call symput('ncate',left(ncate));run;
%end;
%if (&listaconti ne and &listaclass ne) %then%do;
data &archivo.2 (drop=&listaclass &listaconti i);
array &prefijoclass{&ncate} $;
array &prefijoconti{&nconti};
array variclass{&ncate} $ &listaclass;
array variconti{&nconti} &listaconti;
set &archivo;
do i=1 to &nconti;
      &prefijoconti{i}=variconti{i};
end;
do i=1 to &ncate;
      &prefijoclass{i}=variclass{i};
end;run;

data diccionario (keep=original nueva);
do i=1 to &ncate;
cosa="&listaclass";original=scanq(cosa,i);
nueva=cats("&prefijoclass",i);
output;
end;
run;%end;
%else%if (&listaconti eq and &listaclass ne) %then%do;
data &archivo.2 (drop=&listaclass i);
array &prefijoclass{&ncate} $;
array variclass{&ncate} $ &listaclass;
set &archivo;
do i=1 to &ncate;
      &prefijoclass{i}=variclass{i};end;run;%end;
%else%if (&listaconti ne and &listaclass eq) %then%do;
data &archivo.2 (drop=&listaconti i);
array &prefijoconti{&nconti};
array variconti{&nconti} &listaconti;
set &archivo;
do i=1 to &nconti;
      &prefijoconti{i}=variconti{i};
end;run;%end;%mend;
/*------------------------paso 3 INTERACCIONES----------------------*/
%interacttodo(archivo=base,vardep=precio,
listclass=H   OFERTA   FECHA   ESTRELLAS   FECHAINSC   ZONA   CODPOSTAL
COMENTARIOS USABLE ESTACIONAL DIA FINDE,
listconti=CALIDADPRECIO    CONFORT    INSTALACIONESSERVICIOS    LIMPIEZA
PERSONAL UBICACION VALGLOBAL WIFI,
interac=1,directorio= C:\Users\Inma Gutiérrez\Documents\MÁSTER\TFM);
%macro
interacttodo(archivo=,vardep=,listclass=,listconti=,interac=1,director
io=c:);
proc printto print="&directorio\kaka.txt";run;
data     _null_;file     "&directorio\inteconti.txt";put     '     ';file
"&directorio\intecategor.txt";put ' ';run;
data _null_;
length clase conti con cruce1 $ 32000 cruce2 $ 32000;
clase="&listclass";
conti="&listconti";
   ncate= 1;
```

```
   do while (scan(clase, ncate) ^= '');
      ncate+1;
   end;
   ncate+(-1);
   put ncate=;
   nconti= 1;
   do while (scan(conti, nconti) ^= '');
      nconti+1;
   end;
   nconti+(-1);
   put nconti=;

call symput('ncate',left(ncate));
call symput('nconti',left(nconti));

%if&interac=1%then%do;
cruce2=' ';
do i=1 to ncate;
      do j=1 to nconti;
      ca=scan(clase,i);
            con=scan(conti,j);
            cruce1=cats(ca,'*',con);
            file "&directorio\inteconti.txt" mod;
            put cruce1;
      end;
end;

cruce2=' ';
do i=1 to ncate-1;
      do j=i+1 to ncate;
      ca=scan(clase,i);
            con=scan(clase,j);
            if i ne j then cruce1=cats(ca,'*',con);else cruce1=' ';
            file "&directorio\intecategor.txt" mod;
            put cruce1;
      end;
end;run;%end;
data union;run;

/* variables de clase solas */
%if&listclass ne %then%do i=1%to&ncate;
data _null_;cosa="&listclass";va=scanq(cosa,&i);
call symput ('vari',va);run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &vari;
model &vardep=&vari /selection=none;run;

proc print data=tanova;run;

data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep AIC;run;
data b(keep=Fvalue probf);set tanova;if _n_=1 then output;stop;run;
data c;length variable $ 1000;merge a b;variable="&vari";run;
data union;set union c;run;
%end

/* interacciones de variables de clase */

%if&interac=1%then%do;
%if&ncate>1%then%do;
```

```
data pr234;
length vari $ 1000;
infile "&directorio\intecategor.txt";
input vari;
run;
data _null_;set pr234 nobs=nume;ko=nume;
call symput('nintecat',left(ko));stop;run;

%if&listclass ne %then%do i=1%to&nintecat;
data _null_;ko=&i;
set pr234 point=ko;
var1=scan(vari,1);
var2=scan(vari,2);
lista1=compbl(var1||'  '||var2);
call symput('lista1',left(lista1));
call symput('vari',left(vari));
stop;run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &lista1;
model &vardep=&vari / selection=none;run;

data a;set ajuste (where=(Label1='AIC'));
AIC=cvalue1;keep AIC;
data b(keep=Fvalue probf);set tanova;if _n_=1 then output;stop;run;
data c;length variable $ 1000;merge a b;variable="&vari";run;
data union;set union c;run;
%end;
data _null_;if _n_=1 then put 'LISTA CLASE E INTERACCIONES';set
union;put variable @@;run;
%end; %end;

/* variables continuas solas */
%if&listconti ne %then%do i=1%to&nconti;
data _null_;cosa="&listconti";va=scanq(cosa,&i);
call symput ('vari',va);run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
model &vardep=&vari /selection=none;run;
data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep AIC;run;
data b(keep=Fvalue probf);set tanova;if _n_=1 then output;stop;run;
data c;length variable $ 1000;merge a b;variable="&vari";run;
data union;set union c;run;%end;

/* interacciones de variables de clase con variables continuas */
%if&interac=1%then%do;
data pr235;
length vari $ 1000;
infile "&directorio\inteconti.txt";
input vari;run;

data _null_;set pr235 nobs=nume;ko=nume;
call symput('ninteconti',left(ko));stop;run;

%if (&listclass ne) and (&listconti ne) %then%do i=1%to&ninteconti;
data _null_;ko=&i;
set pr235 point=ko;
var1=scan(vari,1);
```

```
var2=scan(vari,2);
call symput('lista1con',left(var1));
call symput('varicon',left(vari));
stop;run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &lista1con;
model &vardep=&varicon / selection=none;run;

data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep AIC;
data b(keep=Fvalue probf);set tanova;if _n_=1 then output;stop;run;
data c;length variable $ 1000;merge a b;variable="&varicon";run;
data union;set union c;run;%end;%end;
proc printto;run;
data union;set union;if _n_=1 then delete;run;
proc sort data=union;by AIC;
proc print data=union;run;
data _null_;set union;put variable @@;run;
%mend;
*---------4 PARTE ESTUDIO DE LA ESTABILIDAD DEL MODELO--------S*/
%randomselect(data=BASE,
listclass=H OFERTA FECHA,
vardepen=PRECIO,
modelo=CONFORT INSTALACIONESSERVICIOS PERSONAL UBICACION VALGLOBAL
WIFI H OFERTA FECHA,
criterio=SBC,
sinicio=12345,
sfinal=12355,
fracciontrain=0.68,
directorio=C:\Users\Inma Gutiérrez\Documents\MÁSTER\TFM);

/*------5 PARTE TRANSFORMACIONES DE VARIABLES----------*/
data TransBase (drop=i);
array x(8) factor1 calidadprecio ubicacion valglobal cpprom fechaprom
ofertaprom estrellasprom; *hay que poner array x(5) porque tenemos 5
variables a transformar;
array z(8);
set discoc.todo;
do i=1to8; z{i}=log(x{i}+1);end;run;

title'LOG(X+1)';
odshtml;
odsgraphicson;
procgplotdata=TRANSbase;
plot PRECIO*z1 PRECIO*z2 PRECIO*z3 PRECIO*z4 PRECIO*z5 PRECIO*z6
PRECIO*z7 PRECIO*z8;run;
%macro
randomselect(data=,listclass=,vardepen=,modelo=,criterio=,sinicio=,sfi
nal=,fracciontrain=,directorio=&directorio);
options nocenter linesize=256;
proc printto print="&directorio\kk.txt";run;
data _null_;file "&directorio\cosa.txt" linesize=2000;run;
%do semilla=&sinicio %to&sfinal;
proc surveyselect data=&data rate=&fracciontrain out=sal1234
seed=&semilla;run;
ods output SelectionSummary=modelos;
ods output SelectedEffects=efectos;
ods output Glmselect.SelectedModel.FitStatistics=ajuste;
proc glmselect data=sal1234 plots=all seed=&semilla;
   class &listclass;
```

```
   model   &vardepen=   &modelo/   selection=stepwise(select=&criterio
   choose=&criterio) details=all stats=all;run;
ods graphics off;
ods html close;
data union;i=5;set efectos;set ajuste point=i;run;
data     _null_;semilla=&semilla;file     "&directorio\cosa.txt"     mod
linesize=2000;set union;put effects ;run;
%end;
proc printto ;run;
data todos;
infile "&directorio\cosa.txt" linesize=2000;
length efecto $ 1000;
input efecto @@;
if efecto ne 'Intercept' then output;run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;

data todos;
infile "&directorio\cosa.txt" linesize=2000;
length efecto $ 1000;
input efecto $ &&;run;
proc freq data=todos;tables efecto /out=salefec;run;
proc sort data=salefec;by descending count;
proc print data=salefec;run;
data _null_;set salefec;put efecto;run;%mend;
/*-----REDES NEURONALES------*/
PROCDMDB  DATA=discoc.todo  dmdbcat=cataprueba;*creamos  un  archivo  de
catalogo;
target precio; * variable dependiente;
var  PRECIO  CALIDADPRECIO  CONFORT  INSTALACIONESSERVICIOS  LIMPIEZA
PERSONAL  UBICACION  VALGLOBAL  WIFI  CPPROM  estrellasprom  zonaprom
diaprom ofertaprom fechaprom factor1 factor2 factor3
; *variables continuas + dependiente;
class OFERTA FECHA USABLE ESTACIONAL FINDE h; run;

DATA DISCOC.SALPREDITESTEARLY;SET SALPREDITEST;RUN;
procneural           data=discoc.trainTODO           dmdbcat=cataprueba
validata=discoc.validaTODO testdata=DISCOC.testOTO;
input  Factor1  CALIDADPRECIO  UBICACION  VALGLOBAL  CPPROM  fechaprom
OFERTAPROM estrellasprom;
target  precio;
hidden 17/ACT=TAnH;

procneural           data=discoc.trainTODO           dmdbcat=cataprueba
validata=discoc.validaTODO testdata=DISCOC.testOTO;
input  Factor1  CALIDADPRECIO  UBICACION  VALGLOBAL  CPPROM  fechaprom
OFERTAPROM estrellasprom;
target  precio;
hidden 17/ACT=TAnH;
/*nloptions maxiter=10000;
netoptions randist=normal ranscale=0.01 random=12345;*/
 train maxiter=70 tech=levmar outest=SALIDA estiter=1;
score data=DISCOC.TRAINTODO out=salpredi outfit=salfit;
score data=DISCOC.VALIDATODO out=salpredival outfit=salfitval;
score data=DISCOC.testOTO out=salpreditest outfit=salfittes;  run;
*grafico early stopping;
symbol1i=join v=circle c=red;
symbol2i=join v=circle c=blue;
procgplotdata=cosa;plot        _AVERR_*_iter_=1        _VAVERR_*_iter_=2
/overlay;run;
```

```
%macrorepito;
data union1;run;
data union2;run;
/*data union3;run; */
%do nodos=1%to30%by1;
proc neural data=DISCOC.TRAINtodo dmdbcat=cataprueba
validata=DISCOC.VALIDAtodo;
input Factor1 CALIDADPRECIO UBICACION VALGLOBAL CPPROM fechaprom
ofertaprom estrellasprom;
target precio;
hidden &nodos/act=tan;
train maxiter=80 tech=levmar;
score data=DISCOC.TRAINtodo out=salpredi outfit=salfit;
score data=DISCOC.VALIDAtodo out=salpredival outfit=salfitval;
/*score data=testcambios out=salpreditest outfit=salfittes; */
run;
data salfit;set salfit;nodos=&nodos;if _n_=2 then output;
data salfitval;set salfitval;nodos=&nodos;if _n_=2 then output;
/*data salfittes;set salfittes;nodos=&nodos;if _n_=2 then output; */
data union1;set union1 salfit;run;
data union2;set union2 salfitval;run;
/*data union3;set union3 salfittes;run; */
%end;
data union1;set union1;
if _n_=1 then delete;run; proc print data=union1;run;
data union2;set union2;
if _n_=1 then delete;run; proc print data=union2;run;
/*data union3;set union3;
if _n_=1 then delete;run; proc print data=union3;run; */
%mend;
%macroactivalcruza;
%let lista='TANH LOG ARC LIN SIN SOF GAU';
%let nume=7;
%do i=1%to&nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
%cruzadaneural(archivo=uno,vardepen=y,conti=x
z,categor=clase,ngrupos=3,sinicio=12345,sfinal=12347,ocultos=10,acti=&
activa);
data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1%to&nume; final&i %end;
%mend;

/*--------------------RANDOM FOREST----------------------*/

%macro randomforest(vardep=,listconti=,listcategor=,maxtrees=,
porcenbag=,numvariables=,maxbranch=,tamhoja=,maxdepth=,pvalor=);

ods listing close;
proc hpforest data=discoc.forest
maxtrees=&maxtrees
trainfraction=0.68
leafsize=&tamhoja
maxdepth=&maxdepth
alpha=&pvalor
exhaustive=5000
missing=useinsearch ;
target precio/level=interval;
input VALGLOBAL LIMPIEZA CONFORT
```

```
UBICACION      INSTALACIONESSERVICIOS      PERSONAL      CALIDADPRECIO
WIFI/level=interval;
input FECHAINSC OFERTA FECHA ESTRELLAS ZONA CODPOSTAL COMENTARIOS DIA
FINDE/level=nominal;
ods          output          fitstatistics=discoc.forest_fitstatistics
variableimportance=discoc.forest_variableimportance;
score out=discoc.salpredi_forest;
run;

data discoc.salpredi_forest; set discoc.salpredi_forest; if (precio=.)
then output; run;
data discoc.salpredi_forest; set discoc.salpredi_forest; id=_N_;run;
data      discoc.salpredi_forest      (keep=P_PRECIO      id);      set
discoc.salpredi_forest;run;
data   discoc.salpredi_forest   (keep=precio   P_PRECIO   id);   merge
discoc.salpredi_forest discoc.valida;by id; run;
data discoc.salpredi_forest; set discoc.salpredi_forest; error=precio-
p_precio; run;
data      discoc.salpredi_forest;      set      discoc.salpredi_forest;
error_cuadrado=error*error; run;

proc   sql   noprint;   create   table   discoc.error   as   select
sum(error_cuadrado) as suma_error_cuad
from discoc.salpredi_forest
;quit;

proc sql noprint; create table Rules as select sum(NRules) as Numrules
from discoc.forest_variableimportance ;run;


            data discoc.criterio; merge discoc.error Rules; run;
            data discoc.criterio; set discoc.criterio;
          Error_cuadratico_medio=suma_error_cuad/1946;/*error   cuadrado
          medio/num observaciones datos valida*/
maxtrees=&maxtrees ;
         trainfraction=0.68;
         leafsize=&tamhoja;
         maxdepth=&maxdepth;
         alpha=&pvalor;
         exhaustive=5000;
             numvariables=&numvariables;
             run;
             data   discoc.resumen_forest;   set   discoc.resumen_forest
discoc.criterio;run;
%mend;
%macroForest;
%let lista2='0.1 0.07 0.05 0.03 0.01';/*pvalor*/
%let nume2=5;
%let lista3='200 150 100 75 50 10 5';/*tam hoja*/
%let nume3=7;
%let lista4='5 10 15 20 25 40 50';/*max arboles*/
%let nume4=6; /*longitud de lista4*/
%do  j=1%to&nume2 %by1;
data                                _null_;p_valor=scanq(&lista2,&j);call
symput('p_valor',left(p_valor));run;
%do  k=1%to&nume3 %by1;
data                                _null_;tamhoja=scanq(&lista3,&k);call
symput('tamhoja',left(tamhoja));run;
%do  l=1%to&nume4 %by1;
data                               _null_;maxtrees=scanq(&lista4,&l);call
symput('maxtrees',left(maxtrees));run;
```

```
      %do  numvariables=3%to15%by3;  /*5*//*cantidad  de  variables  que
usa en cada nodo*/
%do maxdepth=25%to25%by1; /*El maximo es 50*/
%do maxbranch=2%to2%by2; /*5*/

%randomforest(
vardep=precio,
listconti=
FECHAINSC VALGLOBAL LIMPIEZA CONFORT
UBICACION INSTALACIONESSERVICIOS PERSONAL CALIDADPRECIO WIFI,
listcategor=OFERTA  FECHA  ESTRELLAS  ZONA  CODPOSTAL  COMENTARIOS  DIA
FINDE,
maxtrees=&maxtrees.,porcenbag=0.68,numvariables=&numvariables.,maxbran
ch=&maxbranch.,tamhoja=&tamhoja.,maxdepth=&maxdepth.,pvalor=&p_valor.;
%end; %end; %end; %end; %end; %end;%mend;
%forest;
prochpforestdata=discoc.forest_test
maxtrees=20
trainfraction=0.68
leafsize=10
maxdepth=25
alpha=0.07
exhaustive=5000
missing=useinsearch ;
target precio/level=interval;
input VALGLOBAL LIMPIEZA CONFORT
UBICACION      INSTALACIONESSERVICIOS      PERSONAL      CALIDADPRECIO
WIFI/level=interval;
input  FECHAINSC  OFERTA  FECHA  ESTRELLAS  ZONA  CODPOSTAL  COMENTARIOS  DIA
FINDE/level=nominal;
/*ods       output       fitstatistics=discoc.forest_fitstatisticstest
variableimportance=discoc.forest_variableimportancetest;*/
score                                  out=discoc.salpredi_foresttest
outfit=discoc.salpredi_foresttestfit;
run;
/*cajas y bigotes del residuo modelo optimo*/
data     discoc.salpredi_prueba;     set     discoc.salpredi_foresttest;
modelo="optimo";run;
procboxplotdata=discoc.salpredi_prueba;plot R_PRECIO*modelo;run;

/*-------------------------GRADIENT BOOSTING----------------------*/
%macroboosting;
%let lista1='100 200 300 400 500 600'; /*tamhoja*/
%let nume1=6;
%let lista2='125 100 75 50 25 10'; /*maxtrees*/
%let nume2=6;
%do  k=1%to&nume1 %by1;
data                              _null_;tamhoja=scanq(&lista1,&k);call
symput('tamhoja',left(tamhoja));run;
%do  l=1%to&nume2 %by1;
data                             _null_;maxtrees=scanq(&lista2,&l);call
symput('maxtrees',left(maxtrees));run;
%do maxdepth=25%to25%by1; /*El maximo es 50*/
%do maxbranch=2%to2%by2; /*5*/
%let lista3='0.01 0.03 0.05 0.07 0.09 1.1 1.3 1.5 1.7';
%let nume3=9;
%do  m=1%to&nume3 %by1;
data                               _null_;shrink=scanq(&lista3,&m);call
symput('shrink',left(shrink));run;
%let lista4='10 25 50 75 100';
%let nume4=5;
```

```
%do  n=1%to&nume4 %by1;
data                              _null_;iterations=scanq(&lista4,&n);call
symput('iterations',left(iterations));run;

proc        treeboost        data=discoc.train        shrinkage=&shrink
maxbranch=&maxbranch     maxdepth=&maxdepth     iterations=&iterations
leafsize=&tamhoja;
      input   FECHAINSC   OFERTA   FECHA   ESTRELLAS   ZONA   CODPOSTAL
COMENTARIOS DIA FINDE/level=nominal;
      input       VALGLOBAL       LIMPIEZA       CONFORT       UBICACION
INSTALACIONESSERVICIOS PERSONAL CALIDADPRECIO WIFI/level=interval;
      target precio/level=interval;
      SAVE    FIT=discoc.FIT    IMPORTANCE=discoc.IMP    MODEL=discoc.MDL
RULES=discoc.RULES;
      score data=discoc.valida out=discoc.salpredi_boosting;run;
ods listing ;
DATA discoc.salpredi_boosting; SET discoc.salpredi_boosting; run;
data     discoc.salpredi_boosting     (keep=P_PRECIO     PRECIO);     set
discoc.salpredi_boosting;run;
data     discoc.salpredi_boosting;     set     discoc.salpredi_boosting;
error=precio-p_precio; run;
data     discoc.salpredi_boosting;     set     discoc.salpredi_boosting;
error_cuadrado=error*error; run;
proc    sql    noprint;    create    table    discoc.error    as    select
sum(error_cuadrado) as suma_error_cuadra
from discoc.salpredi_boosting;quit;
proc sql noprint; create table Rules as select sum(NRules) as Numrules
from discoc.IMP;run;
            data discoc.criterio; merge discoc.error Rules; run;
            data discoc.criterio; set discoc.criterio;
        Error_cuadratico_medio=suma_error_cuadra/1946;/*error  cuadrado
medio/num observaciones datos valida*/
maxtrees=&maxtrees ;
        leafsize=&tamhoja;
        maxdepth=&maxdepth;
        exhaustive=5000;
            shrinkage=&shrink;
            iterations=&iterations;
            maxbranch=&maxbranch;
            run;
            data  discoc.resumen_boosting;  set  discoc.resumen_boosting
discoc.criterio;run;
        proc sql noprint; create table Rules as select sum(NRules) as
Numrules from discoc.IMP;run;
%end;%end;%end;%end;%end;%end;%mend;

%boosting;
proctreeboost  data=discoc.train  shrinkage=1  maxbranch=4  maxdepth=25
iterations=25 leafsize=100;
      input   FECHAINSC   OFERTA   FECHA   ESTRELLAS   ZONA   CODPOSTAL
COMENTARIOS DIA FINDE/level=nominal;
      input       VALGLOBAL       LIMPIEZA       CONFORT       UBICACION
INSTALACIONESSERVICIOS PERSONAL CALIDADPRECIO WIFI/level=interval;
      target precio/level=interval;
      SAVE    FIT=discoc.FIT    IMPORTANCE=discoc.IMP    MODEL=discoc.MDL
RULES=discoc.RULES;
      score data=discoc.valida out=discoc.salpredi_boostinghojas;run;
```

8. Bibliografía

8.1. Bibliografía referenciada en el texto

1. Breiman, L. (2001). Random Forest. *Machine Learning* , 5-32.
2. Caicedo Bravo, E., & López Sotelo, J. (2009). *Una aproximación práctica a las Redes Neuronales artificiales.* CALI, Programa Editorial Universidad del Valle.
3. (En línea 2016 a). Obtenido de http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/20factor_SPSS.pdf
4. (En línea 2016 b). Obtenido de https://es.wikipedia.org/wiki/Aprendizaje_basado_en_%C3%A1rboles_de_decisi%C3%B3n
5. (En línea 2016 c). Obtenido de https://en.wikipedia.org/wiki/Bootstrap_aggregating
6. Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. En *Annals of Statistics, 1189-1232.*
7. Kiers, H., & Rasson, J.-P. (2000). Cluster Analysis. En *Data Analysis, Classification and Related MEthods.*Belgium: Springer.
8. Peña, D. (2002). *Análisis de Datos Multivariantes.* S.A. MCGRAW-HILL / INTERAMERICANA DE ESPAÑA.
9. Portela, J. (2015). Material Didáctico. *Redes Neuronales, Máster en Minería de Datos e Inteligencia de Negocios* . Madrid.
10. Portela, J. (2015). Material Didáctico. *Asignatura Técnica y Metodología de la Minería de Datos, Máster en Minería de Datos e Inteligenciade Negocios* . Madrid.
11. Rodríguez, & al, e. (2003). METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINNING. *Universidad de Oviedo* . Área de Matemática Aplicada.

8.2. Bibliografía no referenciada en el texto

SAS/STAT 9.2 User's Guide, 2008, Institute Inc., Cary, NC, USA
(En línea) Obtenido de*http://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf*
Prácticas de Estadística y programación en SAS. Llorenç Badiella Busquets, Anna Espinal Berenguer y Joan Valls Marsal.Universidad Autónoma de Barcelona.
(En línea) Obtenido de *http://www.ub.edu/stat/personal/cuadras/metodos.pdf*
(En línea) Obtenido de *http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf*
(En línea) Obtenido de *http://www.uoc.edu/in3/emath/docs/RegresionLineal.pdf*
(En línea) Obtenido de *https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm*
(En línea) Obtenido de *http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema3dm.pdf*
Ronald M. Weiers. *Introduction to Business Statistics.*s.l: Nora Heink, 2011